

# Gegenbauer Collocation Integration Methods: Advances In Computational Optimal Control Theory



Kareem Elgindy  
School of Mathematical Sciences  
Monash University

A Thesis Submitted for the Degree of  
*Doctor of Philosophy*

2013

This page is intentionally left blank

## Copyright Notices

### Notice 1

Under the Copyright Act 1968, this thesis must be used only under the normal conditions of scholarly fair dealing. In particular no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this thesis.

### Notice 2

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

This page is intentionally left blank

I would like to dedicate this thesis to my loving wife Rana and my  
gorgeous daughter Talya...

This page is intentionally left blank

## Acknowledgements

The journey of gaining my Ph.D. degree has been one of the most important learning experiences of my life. As an Egyptian Ph.D. student who used to study among his family, relatives and friends, pursuing Ph.D. studies overseas alone in a very distant country like Australia, and away from all of my beloved people was certainly a hard adventure for me. Yet, they that sow in tears shall reap in joy. Indeed, gaining my Ph.D. degree has provided me with an immense sense of accomplishment. Perhaps what I learned most from this journey is that persistence, self-confidence, and strong belief in one's abilities are the key elements to achieve any hard goals. Moreover, life has taught me to pursue excellence in all that I do. At the end of my three and half years study in Monash University, I feel like I have come a long way in terms of my professional and personal growth. Here I must acknowledge that after the first two years of my candidature, my life became much easier when I found my life partner Rana and married her. Perhaps this was the first time I really felt settlement in Australia. Later Mighty Allah has blessed me with the grace of the birth of my first baby girl Talya. My happiness to gain my Ph.D. degree could not have been completed without the joy of having such a beautiful family. In fact, I cannot stress enough upon the importance of the role played by my wife Rana in gaining my Ph.D. degree as she has always been confident of my capabilities, even at the times when I was in doubt. To her, I would like to say that so much of what I have accomplished is because of you. Thank you so much for being there to support me and standing by me at all times. You played the most instrumental role in my success as a doctoral candidate. Every time I felt like giving up, you were there to encourage me and helped me get through this. You and my baby Talya have been the reason of my pleasure during my Ph.D. long journey. It is to Rana and Talya after Mighty Allah whom I dedicate this dissertation to.

I would like also to thank my supervisors Professor Kate Smith-Miles and Dr. Boris Miller for their valuable comments and suggestions on my work. I thank Professor Kate Smith-Miles also for her hospitality

and introducing me to her very sweet family: Bill, Charlotte, and Jamie. I really enjoyed their company during my stay in Melbourne and I consider them as friends. I deeply thank my examiners, Professor Leslie Greengard and Dr. Ryan Loxton, for providing encouraging and constructive feedback. To them, I would like to say that it is such a great honor for me indeed to have you both as my examiners on my doctoral thesis. I am very grateful for your thoughtful and detailed comments.

I thank my loving parents Taha and Somya, and my sincere sister Noha and brother Islam for their sustained support. I also thank Professor John Lattanzio for his professional advices during my candidature. To him, I would like to say that you have been a true friend of mine. To my fellow members of the School of Mathematical Sciences: Dr. Simon Clarke, Dr. Dianne Atkinson, Dr. Bolis Basit, Dr. Anthony W C Lun, Dr. Kais Hamza, Dr. Jerome Droniou, Dr. Tim Garoni, Mark Bentley, Laura Villanova, Ying Tan, Olivia Mah, Yanfei Kang, and Jieyen Fan, I would like to say thank you for all the good times we had together. I am confident that when I will look back on these days, I will fondly remember each one of you and the role you played in this journey. To the school manager: Mrs. Gertrude Nayak and to the Executive Assistant to Head of School: Mrs. Andrea Peres, I would like to thank you a lot for all the help you have offered to me, especially when I was feeling low. To the IT Service Centre Senior Support Officer: Trent Duncan, your help for handling many hardware and software issues of my computer is greatly appreciated. To my fellow members of the School of Chemistry: Dr. Ayman Nafady and Muhammad Amer, I would like to say that your friendship has been one of my journey's benefits. To all of my other friends: Dr. Salwan Alorfaly and Dr. Jawdat Tashan, I would like to say that you will always be remembered as very dear friends and the time we stayed together was very precious and invaluable. To all of my friends in Australia, thank you for being my family away from my family. Finally, to Melbourne City, Australia, I would like to say that you certainly deserve to be named the world's most liveable city.

## Abstract

The analytic solutions of simple optimal control problems may be found using the classical tools such as the calculus of variations, dynamic programming, or the minimum principle. However, in practice, a closed form expression of the optimal control is difficult or even impossible to determine for general nonlinear optimal control problems. Therefore such intricate optimal control problems must be solved numerically.

The numerical solution of optimal control problems has been the subject of a significant amount of study since the last century; yet determining the optimal control within high precision remains very challenging in many optimal control applications. The classes of direct orthogonal collocation methods and direct pseudospectral methods are some of the most elegant numerical methods for solving nonlinear optimal control problems nowadays. These methods offer several advantages over many other popular discretization methods in the literature. The key idea of these methods is to transcribe the infinite-dimensional continuous-time optimal control problem into a finite-dimensional nonlinear programming problem. These methods are based on spectral collocation methods, which have been extensively applied and actively used in many areas.

Many polynomials approximations and various discretization points have been introduced and studied in the literature for the solution of optimal control problems using control and/or state parameterizations. The commonly used basis polynomials in direct orthogonal collocation methods and direct pseudospectral methods are the Chebyshev and Legendre polynomials, and the collocation points are typically chosen to be of the Gauss or Gauss-Lobatto type of points. The integral operation in the cost functional of an optimal control problem is usually approximated by the well-known Gauss quadrature rules. The differentiation operations are frequently calculated by multiplying a constant differentiation matrix known as the spectral differentiation matrix by the matrix of the function values at a certain discretization/collocation nodes. Thus, the cost functional, the

dynamics, and the constraints of the optimal control problem are approximated by a set of algebraic equations. Unfortunately, there are two salient limitations associated with the applications of typical direct orthogonal collocation methods and direct pseudospectral methods: (i) The spectral differentiation matrix, especially those of higher-orders, are widely known to be ill-conditioned; therefore, the numerical computations may be very sensitive to round-off errors. In fact, for a higher-order spectral differentiation matrix, the ill-conditioning becomes very extreme to the extent that the development of efficient preconditioners is a necessity. (ii) The popular spectral differentiation matrix employed frequently in the literature of direct orthogonal collocation methods and direct pseudospectral methods is a square and dense matrix. Therefore, to determine approximations of higher-orders, one usually has to increase the number of collocation points in a direct pseudospectral method, which in turn increases the number of constraints and the dimensionality of the resulting nonlinear programming problem. Also increasing the number of collocation points in a direct orthogonal collocation method increases the number of constraints of the reduced nonlinear programming problem. Eventually, the increase in the size of the spectral differentiation matrix leads to larger nonlinear programming problems, which may be computationally expensive to solve and time-consuming.

The research goals of this dissertation are to furnish an efficient, accurate, rapid and robust optimal control solver, and to produce a significantly small-scale nonlinear programming problem using considerably few collocation points. To this end, we introduce a direct optimization method based on a novel Gegenbauer collocation integration scheme which draws upon the power of the well-developed nonlinear programming techniques and computer codes, and the well-conditioning of the numerical integration operators. This modern technique adopts two principle elements to achieve the research goals: (i) The discretization of the optimal control problem is carried out within the framework of a complete integration environment to take full advantage of the well-conditioned numerical integral operators. (ii) The integral operations included in the components of the optimal control problem are approximated through a novel optimal numerical quadrature in a certain optimality measure. The introduced numerical quadrature outperforms classical spectral quadratures in accuracy, and can be established efficiently through the Hadamard multiplication of a constant rectangular spectral integration matrix by the vector of the integrand function

values at some optimal Gegenbauer-Gauss interpolation nodes, which usually differ from the employed integration/collocation nodes. The work presented in this dissertation shows clearly that the rectangular form of the developed numerical integration matrix is substantial for the achievement of very precise solutions without affecting the size of the reduced nonlinear programming problem.

Chapter 1 is an introductory chapter highlighting the strengths and the weaknesses of various solution methods for optimal control problems, and provides the motivation for the present work. The chapter concludes with a general framework for using Gegenbauer expansions to solve optimal control problems and an overview for the remainder of the dissertation.

Chapter 2 presents some preliminary mathematical background and basic concepts relevant to the solution of optimal control problems. In particular, the chapter introduces some key concepts of the calculus of variations, optimal control theory, direct optimization methods, Gegenbauer polynomials, Gegenbauer collocation, in addition to some other essential topics.

Chapter 3 presents a published article in Journal of Computational and Applied Mathematics titled “Optimal Gegenbauer quadrature over arbitrary integration nodes.” In this chapter, we introduce a novel optimal Gegenbauer quadrature to efficiently approximate definite integrations numerically. The novel numerical scheme introduces the idea of exploiting the strengths of the Chebyshev, Legendre, and Gegenbauer polynomials through a unified approach, and using a unique numerical quadrature. In particular, the numerical scheme developed employs the Gegenbauer polynomials to achieve rapid rates of convergence of the quadrature for the small range of the spectral expansion terms. For a large-scale number of expansion terms, the numerical quadrature has the advantage of converging to the optimal Chebyshev and Legendre quadratures in the  $L^\infty$ -norm and  $L^2$ -norm, respectively. The developed Gegenbauer quadrature can be applied for approximating integrals with any arbitrary sets of integration nodes. Moreover, exact integrations are obtained for polynomials of any arbitrary degree  $n$  if the number of columns in the developed Gegenbauer integration matrix is greater than or equal to  $n$ . The error formula for the Gegenbauer quadrature is derived. Moreover, a study on the error bounds and the convergence rate shows that the optimal Gegenbauer quadrature exhibits very rapid convergence rates faster than any finite power of the number of Gegenbauer ex-

pansion terms. Two efficient computational algorithms are presented for optimally constructing the Gegenbauer quadrature, and to ideally maintain the robustness and the rapid convergence of the discrete approximations. We illustrate the high-order approximations of the optimal Gegenbauer quadrature through extensive numerical experiments including comparisons with conventional Chebyshev, Legendre, and Gegenbauer polynomial expansion methods. The present method is broadly applicable and represents a strong addition to the arsenal of numerical quadrature methods.

Chapter 4 presents a published article in *Advances in Computational Mathematics* titled “On the optimization of Gegenbauer operational matrix of integration.” The chapter is focused on the intriguing question of “which value of the Gegenbauer parameter  $\alpha$  is optimal for a Gegenbauer integration matrix to best approximate the solution of various dynamical systems and optimal control problems?” The chapter highlights those methods presented in the literature which recast the aforementioned problems into unconstrained/constrained optimization problems, and then add the Gegenbauer parameter  $\alpha$  associated with the Gegenbauer polynomials as an extra unknown variable to be optimized. The theoretical arguments presented in this chapter prove that this naive policy is invalid since it violates the discrete Gegenbauer orthonormality relation, and may in turn produce false optimization problems analogs to the original problems with poor solution approximations.

Chapter 5 presents a published article in *Journal of Computational and Applied Mathematics* titled “Solving boundary value problems, integral, and integro-differential equations using Gegenbauer integration matrices.” The chapter resolves the issues raised in the previous chapter through the introduction of a hybrid Gegenbauer collocation integration method for solving various dynamical systems such as boundary value problems, integral and integro-differential equations. The proposed method recasts the original problems into their integral formulations, which are then discretized into linear systems of algebraic equations using a hybridization of the Gegenbauer integration matrices developed in Chapter 3. The resulting linear systems are generally well-conditioned and can be easily solved using standard linear system solvers. A study on the error bounds of the proposed method is presented, and the spectral convergence is proven for two-point boundary-value problems. Comparisons with other competitive methods in the recent literature are included. The proposed method

results in an efficient algorithm, and spectral accuracy is verified using eight test examples addressing the aforementioned classes of problems. The developed numerical scheme provides a viable alternative to other solution methods when high-order approximations are required using only a relatively small number of solution nodes.

Chapter 6 presents a published article in The Proceedings of 2012 Australian Control Conference, AUCC 2012, titled “Solving optimal control problems using a Gegenbauer transcription method.” The chapter presents a novel direct orthogonal collocation method using Gegenbauer-Gauss collocation for solving continuous-time optimal control problems with nonlinear dynamics, state and control constraints, where the admissible controls are continuous functions. The framework of the novel method involves the mapping of the time domain onto the interval  $[0, 1]$ , and transforming the dynamical system given as a system of ordinary differential equations into its integral formulation through direct integration. In this manner, the proposed Gegenbauer transcription method unifies the process of the discretization of the dynamics and the integral cost function. The state and the control variables are then fully parameterized using Gegenbauer expansion series with some unknown Gegenbauer spectral coefficients. The proposed Gegenbauer transcription method recasts the performance index, the reduced dynamical system, and the constraints into systems of algebraic equations using the optimal Gegenbauer quadrature introduced in Chapter 3. Finally, the Gegenbauer transcription method transcribes the infinite-dimensional optimal control problem into a finite-dimensional nonlinear programming problem, which can be solved in the spectral space; thus approximating the state and the control variables along the entire time horizon. The high precision and the spectral convergence of the discrete solutions are verified through two optimal control test problems with nonlinear dynamics and some inequality constraints. In particular, we investigate the application of the proposed method for finding the best path in 2D of an unmanned aerial vehicle moving in a stationary risk environment. Moreover, we compare the performance of the proposed Gegenbauer transcription method with another classical variational technique to demonstrate the efficiency and the accuracy of the proposed method.

Chapter 7 presents a published article in Journal of Computational and Applied Mathematics titled “Fast, accurate, and small-scale direct trajectory optimization using a Gegenbauer transcription method.” This chapter extends the Gegenbauer transcription method intro-

duced in the preceding chapter to deal further with continuous-time optimal control problems including different orders time derivatives of the states by solving the continuous-time optimal control problem directly for the control  $u(t)$  and the highest-order time derivative  $x^{(N)}(t)$ ,  $N \in \mathbb{Z}^+$ . The state vector and its derivatives up to the  $(N - 1)$ th-order derivative can then be stably recovered by successive integration. Moreover, we present our solution method for solving linear quadratic regulator problems as we aim to cover a wider collection of continuous-time optimal control problems with the concrete aim of comparing the efficiency of the current work with other classical discretization methods in the literature. The advantages of the proposed direct Gegenbauer transcription method over other traditional discretization methods are shown through four well-studied optimal control test examples. The present work is a major breakthrough in the area of computational optimal control theory as it delivers significantly accurate solutions using considerably small numbers of collocation points, states and controls expansions terms. Moreover, the Gegenbauer transcription method produces very small-scale nonlinear programming problems, which can be solved very quickly using modern nonlinear programming software. The Gegenbauer collocation integration scheme adopted in this dissertation allows for the solution of continuous-time optimal control problems governed by various types of dynamical systems; thus encompassing a wider collection of problems than standard optimal control solvers. Moreover, the method is simple and very suitable for digital computations.

Chapter 8 presents some concluding remarks on the works developed in this dissertation including some suggestions for future research.

## List of Publications

Elgindy, K. T., Smith-Miles, K. A., 15 October 2013. Fast, accurate, and small-scale direct trajectory optimization using a Gegenbauer transcription method. *Journal of Computational and Applied Mathematics* 251 (0), 93–116.

Elgindy, K. T., Smith-Miles, K. A., April 2013. Optimal Gegenbauer quadrature over arbitrary integration nodes. *Journal of Computational and Applied Mathematics* 242 (0), 82–106.

Elgindy, K. T., Smith-Miles, K. A., 1 January 2013. Solving boundary value problems, integral, and integro-differential equations using Gegenbauer integration matrices. *Journal of Computational and Applied Mathematics* 237 (1), 307–325.

Elgindy, K. T., Smith-Miles, K. A., 1 December 2012. On the optimization of Gegenbauer operational matrix of integration. *Advances in Computational Mathematics*, Springer US, 1–14. DOI 10.1007/s10444-012-9289-5.

Elgindy, K. T., Smith-Miles, K. A., and Miller, B., 15–16 November 2012. Solving optimal control problems using a Gegenbauer transcription method. In: *The Proceedings of 2012 Australian Control Conference, AUCC 2012*. Engineers Australia, University of New South Wales, Sydney, Australia.

This page is intentionally left blank

## PART A: General Declaration

### Monash University

#### Declaration for thesis based or partially based on conjointly published or unpublished work

#### General Declaration

In accordance with Monash University Doctorate Regulation 17 Doctor of Philosophy and Research Master's regulations the following declarations are made:

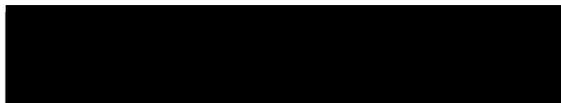
I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 5 original papers published in peer reviewed journals. The core theme of the thesis is the development of accurate, robust and efficient numerical methods for the solution of continuous-time nonlinear optimal control problems. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the candidate, working within the School of Mathematical Sciences, Monash University under the supervision of Professor Kate Smith-Miles and Doctor Boris Miller. The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of Chapters 3-7 my contribution to the work involved the following:

Thesis chapter	Publication title	Publication status*	Nature and extent of candidate's contribution
3	Optimal Gegenbauer quadrature over arbitrary integration nodes	Published	The author of the key ideas, Programming codes, organization, development, and writing up of the article
4	On the optimization of Gegenbauer operational matrix of integration	Published	The author of the key ideas, organization, development, and writing up of the article
5	Solving boundary value problems, integral, and integro-differential equations using Gegenbauer integration matrices	Published	The author of the key ideas, programming codes, organization, development, and writing up of the article
6	Solving optimal control problems using a Gegenbauer transcription method	Published	The author of the key ideas, programming codes, organization, development, and writing up of the article
7	Fast, accurate, and small-scale direct trajectory optimization using a Gegenbauer transcription method	Published	The author of the key ideas, programming codes, organization, development, and writing up of the article

I have renumbered sections of published papers in order to generate a consistent presentation within the thesis.



**Signed:**

**Date:** 10/05/2013.....

This page is intentionally left blank

# Contents

<b>Contents</b>	<b>xviii</b>
<b>List of Figures</b>	<b>xxii</b>
<b>List of Tables</b>	<b>xxviii</b>
<b>List of Acronyms</b>	<b>xxx</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Optimal Control (OC)	2
1.1.1 Historical Background and Applications	2
1.1.2 Elements of OC Problems	3
1.1.3 Analytical Methods and Classes of Numerical Methods	3
1.1.4 Advantages of DOMs	6
1.1.5 Elements of DOMs and the Shooting Methods	7
1.1.6 Typical Discretization Methods for Dynamical Systems	8
1.2 Spectral Methods	8
1.2.1 Orthogonal Collocation/Pseudospectral (PS) Methods	9
1.2.2 Direct Orthogonal Collocation Methods (DOCMs) and Direct PS Methods	10
1.2.3 DOCMs Versus Shooting Methods and IOMs	11
1.2.4 DOCMs Versus Standard Direct Local Collocation Methods (DLCMs)	11
1.2.5 Orthogonal Functions	12
1.2.6 Jacobi Polynomials	13
1.3 Motivation of the Present Work	13
1.3.1 Collocation Points	17
1.3.2 Choice of the NLP Solver	18
1.3.3 Significance of the Operational Matrix of Integration	18
1.4 Framework	19
1.4.1 Advantages of the GTM	20

1.5	Thesis Overview . . . . .	22
<b>2</b>	<b>Preliminary Mathematical Background</b>	<b>28</b>
2.1	CV . . . . .	28
2.2	OC Theory . . . . .	29
2.2.1	Formulation of CTOCPs . . . . .	30
2.2.1.1	Case with Fixed Final Time and No Terminal or Path Constraints . . . . .	31
2.2.1.2	The Necessary Optimality Conditions Using the Variational Approach . . . . .	32
2.2.1.3	Case with Terminal Constraints . . . . .	33
2.2.1.4	Case with Input Constraints– The MP . . . . .	33
2.2.1.5	Special Cases– Some Results Due to the MP . . . . .	35
2.2.1.6	General Mathematical Formulation of CTOCPs . . . . .	36
2.2.2	DP– Sufficient Conditions of Optimality . . . . .	36
2.2.2.1	The Curse of Dimensionality– A Fatal Drawback Associated with the DP Solution Method . . . . .	38
2.2.3	DOMs . . . . .	38
2.2.3.1	DCMs . . . . .	41
2.2.3.2	DLCMs . . . . .	43
2.2.3.3	DGCMs . . . . .	43
2.2.3.4	Choice of the Collocation Points in a DOCM/PS Method . . . . .	44
2.2.3.5	The Framework of Solving OC Problems Using DOCMs/PS Methods . . . . .	45
2.2.3.6	DGCMs Versus DLCMs . . . . .	46
2.3	Gegenbauer Polynomials . . . . .	47
2.4	The Gegenbauer Approximation of Functions . . . . .	50
2.5	Gegenbauer Collocation Methods: Convergence Rate . . . . .	51
2.6	The Gegenbauer Operational Matrix of Integration . . . . .	53
2.7	Solving Various Dynamical Systems and OC Problems by Opti- mizing the GIM . . . . .	56
<b>3</b>	<b>Optimal Gegenbauer Quadrature Over Arbitrary Integration Nodes</b>	<b>61</b>
3.1	Introduction . . . . .	61
3.2	Generation of Optimal GIMs . . . . .	64
3.2.1	The Proposed Method . . . . .	67
3.2.2	Generation of the P-matrix and Error Analysis . . . . .	69
3.2.3	Polynomial Integration . . . . .	71
3.2.4	Error Bounds and Convergence Rate . . . . .	72

3.2.5	Convergence of the PMQ to the Chebyshev Quadrature in the $L^\infty$ -norm . . . . .	73
3.2.6	Determining the Interval of Uncertainty for the Optimal Gegenbauer Parameters of the P-matrix for Small/Medium Range Expansions . . . . .	74
3.2.7	Substantial Advantages of the Gegenbauer Collocation Methods Endowed with the PMQ . . . . .	77
3.2.8	The Matrix Form Gegenbauer Approximation of Definite Integrations . . . . .	78
3.2.9	Computational Algorithms . . . . .	79
3.2.9.1	The Nonsymmetric Integration Points Case . . . . .	81
3.2.9.2	The Symmetric Integration Points Case: A More Computationally Efficient Algorithm . . . . .	82
3.3	Numerical Results . . . . .	83
3.3.1	Comparisons with the Optimal QMQ . . . . .	83
3.3.2	Comparisons with the CC Method . . . . .	88
3.4	Further Applications . . . . .	90
3.5	Future Work . . . . .	91
3.6	Conclusion . . . . .	92
3.A	Some Properties of the Gegenbauer Polynomials . . . . .	93
3.B	Proof of Theorem 3.2.3 . . . . .	95
3.C	Proof of Theorem 3.2.7 . . . . .	96
3.D	Proof of Theorem 3.2.8 . . . . .	96
3.E	Proof of Lemma 3.2.9 . . . . .	98
3.F	Proof of Theorem 3.2.10 . . . . .	99
3.G	Algorithm 2.1 . . . . .	101
3.H	Algorithm 2.2 . . . . .	102
<b>4</b>	<b>On the Optimization of Gegenbauer Operational Matrix of Integration</b> . . . . .	<b>107</b>
4.1	Introduction . . . . .	107
4.2	Preliminary Definitions and Properties . . . . .	109
4.3	Solving Various Mathematical Problems by Optimizing the Gegenbauer Integration Matrix . . . . .	112
4.4	The Mathematical Proof . . . . .	114
4.5	Concluding Remarks and a Practical Alternative Method . . . . .	117
<b>5</b>	<b>Solving Boundary Value Problems, Integral, and Integro-Differential Equations Using Gegenbauer Integration Matrices</b> . . . . .	<b>125</b>
5.1	Introduction . . . . .	125
5.2	The GIMs . . . . .	129

5.2.1	The Proposed HGIM . . . . .	134
5.2.2	Convergence Analysis and Error Bounds . . . . .	136
5.3	Numerical Results . . . . .	141
5.4	Concluding Remarks . . . . .	156
<b>6</b>	<b>Solving Optimal Control Problems Using a Gegenbauer Transcription Method</b>	<b>163</b>
6.1	Introduction . . . . .	163
6.2	The OC Problem Statement . . . . .	165
6.3	The GTM . . . . .	165
6.4	Illustrative Numerical Examples . . . . .	168
6.5	Discussion and Conclusion . . . . .	173
6.5.1	Elementary Properties and Definitions . . . . .	176
6.5.2	The Optimal Gegenbauer Quadrature and Definite Integrals Approximations . . . . .	177
<b>7</b>	<b>Fast, Accurate, and Small-Scale Direct Trajectory Optimization Using a Gegenbauer Transcription Method</b>	<b>184</b>
7.1	Introduction . . . . .	184
7.2	The CTOCPs statements . . . . .	188
7.3	The GTM . . . . .	189
7.3.1	Solving Problem $\mathcal{P}_1$ using the GTM . . . . .	191
7.3.2	Solving Problem $\mathcal{P}_2$ using the GTM . . . . .	193
7.4	Properties of the GTM . . . . .	196
7.5	Illustrative Numerical Examples . . . . .	197
7.6	Conclusion and Future Research . . . . .	217
7.A	Elementary Properties and Definitions . . . . .	220
<b>8</b>	<b>Conclusion and Future Research</b>	<b>224</b>
8.1	Summary of Contributions . . . . .	224
8.2	Future Research Directions . . . . .	230
	<b>References</b>	<b>233</b>

# List of Figures

1.1	SDMs are known to be severely ill-conditioned, and their implementation causes degradation of the observed precision. Moreover, it has been shown that the time step restrictions can be more severe than those predicted by the standard stability theory (Trefethen, 1988; Trefethen and Trummer, 1987). For higher-order SDMs, the development of efficient preconditioners is extremely crucial (Elbarbary, 2006; Hesthaven, 1998). . . . .	14
1.2	Any increase in the size of the square SDM in a direct PS method requires the same increase in the number of collocation points, which leads to an increase in the dimension of the NLP problem, and the number of constraints. Eventually, these elements accumulate to yield a larger NLP problem. . . . .	15
1.3	The optimal GIM captures the most suitable properties of the Chebyshev, Legendre, and Gegenbauer polynomials required for a given problem. It is a well-conditioned operator, and its well-conditioning is essentially unaffected for increasing number of grid points. The use of integration for constructing the spectral approximations improves the rate of convergence of the spectral interpolants, and allows the multiple boundary conditions to be incorporated more efficiently. . . . .	20
1.4	The GTM for a continuous-time Bolza OC problem. . . . .	21
2.1	The elements of CTOCPs . . . . .	31

3.1	The steps for evaluating the definite integral $\int_{-1}^{x_i} f(x)dx$ of a given function $f(x) \in C^\infty[-1, 1]$ by using Theorem 3.2.3. The figure shows that instead of strictly using the set of integration nodes $\{x_i\}_{i=0}^N = S_N^{(\alpha)}$ that is the same as the set of interpolation points required for constructing the Gegenbauer quadrature, one chooses any arbitrary set of integration nodes $\{x_i\}_{i=0}^N$ . For a particular integration node $x_i$ , the PMQ determines the optimal Gegenbauer parameter $\alpha_i^*$ in the sense of minimizing the square of the $\eta$ -function, $\eta_{i,M}^2(\alpha)$ . The PMQ then employs the adjoint GG nodes $z_{i,k}$ corresponding to the integration node $x_i$ as the optimal set of interpolation points, and evaluates the integrand $f(x)$ at these optimal points. The Gegenbauer quadrature method proceeds by constructing the $i$ th row of the P-matrix, $(p_{i,0}^{(1)}(\alpha_i^*), p_{i,1}^{(1)}(\alpha_i^*), \dots, p_{i,M}^{(1)}(\alpha_i^*))$ , and evaluates the definite integral $\int_{-1}^{x_i} f(x)dx$ as stated by Formula (3.14). . . . .	71
3.2	The profile of the Gegenbauer weight function $w^{(\alpha)}(x)$ for $\alpha = -0.4, 0, 0.5, 1, 2; 100$ . Clearly the weight function dies out near the boundaries $x = \pm 1$ , and the function becomes nonzero only on a small subdomain centered at the middle of the interval $x = 0$ for increasing values of $\alpha$ . . . . .	76
3.3	The values of $\alpha_i^*$ versus the four sets of integration points $S_{12}^{(0)}, S_{12}^{(\alpha^*)}, S^{3,12}; S^{4,12}$ using $\varepsilon = 0.016, 0.028; 0.01$ . . . . .	86
4	$\text{Log}(EE)$ in floating-point arithmetic for the PMQ and the CC quadrature versus $N = 2, 4, \dots, 20$ , for the seven test functions $\{f_i\}_{i=1}^7$ on $[0, 1]$ . The results of the CC method are reported at $S_N^{(0)}$ . The results of the PMQ are reported at $S_N^{(0)}, S^{3,N}; S^{4,N}$ . . . . .	100
5.1	The error factor $d_N^{(\alpha)}$ decays exponentially fast for increasing values of $N$ . . . . .	141
5.2	The numerical experiments of the HGIM on Example 5.3.1. Figure (a) shows the graph of $y(x)$ on $[0, 1]$ . Figure (b) shows the MSEs of the HGIM for $N = 16; 64$ . Figure (c) shows the MAEs of the HGIM for $N = 7, 15, 23; 31$ . . . . .	143
5.3	The numerical experiments of the HGIM on Example 5.3.2. Figure (a) shows the graph of $y(z)$ on $[0, 1]$ . Figure (b) shows the MSEs for $N = 16, 64; 128$ . . . . .	145
5.4	The numerical experiments of the HGIM on Example 5.3.3. Figure (a) shows the graph of $y(x)$ on $[0, 1]$ . Figure (b) shows the MSEs for $N = 16; 64$ . . . . .	146

5.5	The numerical experiments of the HGIM on Example 5.3.4. Figure (a) shows the graph of $y(x)$ on $[-1, 1]$ . Figure (b) shows the MAEs for $N = 8; 10$ . . . . .	148
5.6	The numerical experiments of the HGIM on Example 5.3.5. Figure (a) shows the graph of $y(x)$ on $[0, 1]$ . Figure (b) shows the MAEs of the HGIM for $N = 8, 16; 32$ . . . . .	150
5.7	The numerical experiments of the HGIM on Example 5.3.6. Figure (a) shows the graph of $y(x)$ on $[-1, 1]$ . Figure (b) shows the MAEs of the HGIM for $N = 3, 5, 7; 9$ . . . . .	153
5.8	The numerical experiments of the HGIM on Example 5.3.7. Figure (a) shows the graph of $y(x)$ on $[0, 1]$ . Figure (b) shows the MAEs of the HGIM for $M = 14; N = 3, 4, 6, 7, 9; 10$ . . . . .	154
5.9	The numerical experiments of the HGIM on Example 5.3.8. Figure (a) shows the graph of $y(x)$ on $[-1, 1]$ . Figure (b) shows the MAE of the Gegenbauer spectral method for $M = 16$ , and $N = 3, 7, 11; 15$ .156	
6.1	The numerical experiments of the GTM on Example 6.4.1. Figure (a) shows the profile of the control history on the calculated flight time domain $[0, 7.188]$ . Figure (b) shows the 2D state trajectory in the hazard relief contour. The results are obtained at $S_8^{(0.2)}$ using $L = M = 8$ . . . . .	172
6.2	The figure shows the projected state trajectory in a black solid line along the 3D hazard relief. $A'; B'$ denote the points $(-3, -4, f(-3, -4)); (3, 3, f(3, 3))$ , respectively. The results are obtained at $S_8^{(0.2)}$ using $L = M = 8$ . . . . .	173
6.3	The numerical experiments of the GTM on Example 6.4.2. Figure (a) shows the profile of the control history on the calculated flight time domain $[0, 6.268]$ . Figure (b) shows the 2D state trajectory in the hazard relief contour. The results are obtained at $S_5^{(-0.1)}$ using $L = M = 5$ . . . . .	174
6.4	The figure shows the projected state trajectory in a black solid line along the 3D hazard relief. $A'; B'$ denote the points $(-3, -4, f(-3, -4)); (3, 3, f(3, 3))$ , respectively. The results are obtained at $S_5^{(-0.1)}$ using $L = M = 5$ . . . . .	175
7.1	The numerical experiments of the GTM on Example 7.5.1. The figure shows the profiles of the exact control and state variables on $[0, 1]$ together with the approximate optimal state and control variables obtained by the GTM. The results are reported at $S_5^{(0.5)}$ using $L = M = M_P = 5$ . . . . .	201

7.2	The sketch of the absolute errors $E_x(t); E_u(t)$ obtained using Gegenbauer collocation at $S_5^{(0.5)}$ with $L = M = 5; M_P = 20$ . Figures (a) & (b) show the values of the absolute errors $E_x; E_u$ at the GG collocation nodes while Figures (c) & (d) show the profiles of the absolute errors on the time interval $[0, 1]$ . It can be clearly seen from the former two figures that the absolute errors are small at the GG points as expected. The latter two figures show that the absolute errors of the state and the control variables are also small over the whole time horizon with $\max_{t \in [0,1]} E_x(t) \approx 6.14412 \times 10^{-6}; \max_{t \in [0,1]} E_u(t) \approx 7.34135 \times 10^{-6}$ , respectively. . . . .	203
7.3	The numerical experiments of the GTM on Example 7.5.2. Figures (a) and (b) show the graphs of the approximate optimal state and control variables on $[0, 1]$ . The results of the GTM are obtained at $S_8^{(0.2)}$ using $L = M = 6$ . The solid, dotted, and dash-dotted lines are generated using 100 nodes. . . . .	206
7.4	The plots of the error function $\mathcal{E}_2(t)$ and the inequality constraint function $\mathcal{E}_3(t)$ produced by the present GTM through Gegenbauer collocation at $S_8^{(0.3)}$ with $L = M = 8$ . Figure (a) shows the magnitude of the error function $\mathcal{E}_2(t)$ at the GG collocation nodes $t_i \in S_8^{(0.3)}$ . Figure (b) shows the propagation of the error function $\mathcal{E}_2(t)$ on the time interval $[0, 1]$ , where it can be clearly seen that the error function is an oscillatory function of small magnitude over the whole time horizon with $\max_{t \in [0,1]}  \mathcal{E}_2(t)  \approx 2.479 \times 10^{-4}$ . Figure (c) shows the profile of the nonnegative inequality constraint function $\mathcal{E}_3(t)$ on the time interval $[0, 1]$ , where it can be verified that the optimal $x_2$ -trajectory never cross the boundary constraint $r(t)$ . . . . .	209
7.5	The numerical experiments of the GTM on Example 7.5.3. Figures (a) and (b) show the graphs of the approximate state variables and the control variable on $[0, 1]$ . The results are obtained at $S_8^{(-0.4)}$ using $L = M = 12$ . The solid, dotted, and dash-dotted lines are generated using 100 nodes. . . . .	210
7.6	The numerical experiments of the GTM on Example 7.5.4. Figures (a) and (b) show the profiles of the states and the control variables on $[0, 1]$ , respectively. The results are reported at $S_6^{(1)}$ using $M_P = 20; L = M = 6$ . . . . .	214

7.7 The sketch of the error function  $\mathcal{E}_2(t)$  built using Gegenbauer collocation at  $S_6^{(1)}$  with  $M_P = 20; L = M = 6$ . Figure (a) shows the magnitude of the error at the GG collocation nodes. Figure (b) shows the profile of the error function  $\mathcal{E}_2(t)$  on the time interval  $[0, 1]$ , where it can be clearly seen that the error function is an oscillatory function of small magnitude over the whole time horizon with  $\max_{t \in [0, 1]} |\mathcal{E}_2(t)| \approx 2.1 \times 10^{-5}$ . . . . . 215

This page is intentionally left blank

# List of Tables

3.1	The PMQ versus the optimal QMQ in approximating the definite integrals of $f_1(x)$ for the set of integration nodes $S_N^{(\alpha^*)}$ . . . . .	84
3.2	The PMQ versus the optimal QMQ in approximating the definite integrals of $f_2(x)$ for the set of integration nodes $S_N^{(\alpha^*)}$ . . . . .	84
3.3	The PMQ versus the optimal QMQ in approximating the definite integrals of $f_3(x)$ for the set of integration nodes $S_N^{(\alpha^*)}$ . . . . .	85
3.4	The PMQ versus the optimal QMQ in approximating the definite integrals of $f_1(x)$ for the set of integration nodes $S_N^{(\alpha^*)}$ . The results of the PMQ are reported for increasing values of $M$ . . . . .	88
5.1	Comparison of the present HGIM with Greengard and Rokhlin's method (Greengard and Rokhlin, 1991). $N_T$ refers to the total number of nodes. The results shown are the observed MAEs of both methods. . . . .	144
5.2	Comparison of the present method with Greengard's method (Greengard, 1991). The results shown are the observed MSEs of both methods. . . . .	147
5.3	Comparison of the present method with Elbarbary's Chebyshev pseudospectral integration method (Elbarbary, 2007). The results are the observed MAEs at each collocation node $x_i$ . The results of the present method are reported at $\alpha = 0.6; 1$ for $N = 8; 10$ , respectively. . . . .	149
5.4	Comparison of the present method with Zahra's sixth-order spline method (Zahra, 2011). . . . .	150
5.5	Comparison of the present method with Long et al. (2009). The results are the observed MAEs of both methods. . . . .	152
5.6	Comparison of the present method with Maleknejad and Attary's method (Maleknejad and Attary, 2011). The results are the observed MAEs in both methods. . . . .	155

5.7	Comparison of the present method with El-Kady et al.'s Gegenbauer integration method (El-Kady et al., 2009). $N_T$ denotes the total number of nodes. The results are the observed MAEs in both methods. . . . .	156
6.1	The results of the present GTM for $V = 2$ , and different values of $\alpha, N, L; M$ . $(MAE)_{BC}$ denotes the MAE at the boundary condition (6.10h). . . . .	171
6.2	The results of the present GTM for $V = 2$ , and different values of $\alpha, N, L; M$ . . . . .	172
7.1	The numerical results obtained by different methods for solving Example 7.5.1. DIM refers to the dimension of the NLP problem. The results of the present GTM are obtained at $S_5^{(0.5)}$ . . . . .	200
7.2	The approximate cost function value $J$ of Example 7.5.2 obtained by different methods. The results of the present GTM are obtained at the GG collocation set $S_N^{(\alpha)}$ , for the shown values of $\alpha; N$ using different values of $L; M$ . . . . .	207
7.3	The approximate cost function of Example 7.5.3 obtained by different methods. . . . .	211
7.4	Comparisons between the present GTM using the P-matrix and the methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011); Sirisena and Tan (1974). The reported results of the Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011) methods were obtained using SNOPT (Gill et al., 2005), and are exactly as quoted from Ref. (Ma et al., 2011). . . . .	212
7.5	The CPU time of the present method using the P-matrix with $M_P = 16, 20; L = M = 10$ , versus the methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011). The results of the present method for $N = 64, 128; 256$ were obtained by collocations at the GG sets $S_{64}^{(-0.1)}, S_{128}^{(0.5)}; S_{256}^{(0.3)}$ using $M_P = 16$ and by collocations at the GG sets $S_{64}^{(0.5)}, S_{128}^{(0.6)}; S_{256}^{(0.3)}$ using $M_P = 20$ . The reported CPU times of the Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011) methods are exactly as quoted from Ref. (Ma et al., 2011). . . . .	213

# List of Acronyms

BVP	boundary value problem
CGL	Chebyshev-Gauss-Lobatto
CTOCP	continuous-time optimal control problem
CV	calculus of variations
DCM	direct collocation method
DGCM	direct global collocation method
DLCM	direct local collocation method
DOCM	direct orthogonal collocation method
DOM	direct optimization method
DP	dynamic programming
GG	Gegenbauer-Gauss
GIM	Gegenbauer integration matrix
GTM	Gegenbauer transcription method
HJB	Hamilton-Jacobi-Bellman
IOM	indirect optimization method
IVP	initial value problem
LG	Legendre-Gauss
LGL	Legendre-Gauss-Lobatto
LGR	Legendre-Gauss-Radau

LQR linear quadratic regulator  
MAE maximum absolute error  
MP minimum principle  
MSE mean square error  
NLP nonlinear programming  
OC optimal control  
ODE ordinary differential equation  
PDE partial differential equation  
PS pseudospectral  
SDM spectral differentiation matrix  
SIM spectral integration matrix  
TPBVP two-point boundary-value problem

*Chapter 1*  
*Introduction*

# Chapter 1

## Introduction

### 1.1 Optimal Control (OC)

#### 1.1.1 Historical Background and Applications

OC theory is one of several applications and extensions of the calculus of variations (CV). The main concerns of OC theory are the analysis and the design of control systems (Bryson, 1996). Developed by outstanding scholars like Johann Bernoulli (1667–1748), Isaac Newton (1642–1727), Leonhard Euler (1707–1793), Ludovico Lagrange (1736–1813), Andrien Legendre (1752–1833), Carl Jacobi (1804–1851), William Hamilton (1805–1865), Karl Weierstrass (1815–1897), Adolph Mayer (1839–1907), and Oskar Bolza (1857–1942) in the 17th - 20th centuries, OC theory started its progress, and arose as a distinguished branch of mathematics with a strong base. The significant milestones in the development of the theory were pioneered by Richard Bellman (1920–1984) in 1953, who invented a new view of the Hamilton-Jacobi theory well-known as dynamic programming (DP)—a major breakthrough in the theory of multistage decision processes— and by Lev Pontryagin (1908–1988) and his students in 1956 who enunciated the elegant maximum principle (or minimum principle (MP)); cf. (Boltyanskii et al., 1956), which is considered a centerpiece of OC theory, and extends the CV to handle control/input variable inequality constraints. Therefore, one can mark the 1950s as the decade in which the discipline of OC theory reached its maturity. No doubt that the arrival of the digital computer in the 1950s is the cornerstone in the development of OC theory providing the impetus for the applications of OC theory to many complicated problems. Before that only fairly simple OC problems could be solved. The breadth and the range of the applications of OC theory are certainly part of the establishment of OC theory as an important and rich area of applied mathematics. That is part of why OC theory is considered an “application-oriented” mathematics. Nowadays OC theory

## Chapter 1

---

is strongly utilized in many areas such as biology, medicine, economics, finance, management sciences, aerospace, physics, engineering, bioengineering, agriculture, process control, robotics, armed forces, astronautics, vibration damping, magnetic control, ascent guidance, oil recovery, collision avoidance maneuvers, sewage pumping stations, water flood control, improving the caliber of the crystal products, smoke control in a traffic link tunnel, and a host of other areas; cf. (Apreutesei, 2012; Bassey and Chigbu, 2012; Chakraborty et al., 2011; Deisenberg and Hartl, 2005; Ding and Wang, 2012; Engelhart et al., 2011; Hua et al., 2011; Kang and Bedrossian, 2007; Lashari and Zaman, 2012; Lenhart and Workman, 2007; Okosun et al., 2011; Paengjuntuek et al., 2012; Picart et al., 2011; Salmani and Büskens, 2011; Smith, 2008; Verhelst et al., 2012; Wei et al., 2012; Yang et al., 2012; Zhang and Swinton, 2012; Zheng et al., 2012). OC theory has received considerable attention by researchers and it continues to be an active research area within control theory with broad attention from industry.

### 1.1.2 Elements of OC Problems

OC theory considers the problem of how to control a given system so that it has an optimal behavior in a certain optimality sense. In other words, the principle goal of OC theory is to determine the control which causes a system to meet a set of physical constraints while optimizing some performance criterion (cost function). The dynamical system (plant) describes the evolution of the system's state and how the controls affect it. Dynamical systems can be engineering systems, economic systems, biological systems, and so on. The cost function is a functional of the state and the control, and describes the cost to be minimized or the utility to be maximized. It represents a desired metric such as time, energy costs, fuel consumption, productivity, or any other parameter of interest in a given application. The control which optimizes the cost functional while satisfying a certain set of physical constraints is called the OC. The OC can be described as an optimization device or instrument which influences the system to reach a desired state in minimum-time, or with minimal energy costs, or to achieve maximal productivity in a certain time, etc.

### 1.1.3 Analytical Methods and Classes of Numerical Methods

Theoretically, the OC may be derived using the necessary conditions provided by the CV and the MP, or by solving the sufficient condition described by the Hamilton-Jacobi-Bellman (HJB) equation. However, except for special cases, most OC problems cannot be solved analytically, and a closed form expression for the OC is usually out of reach. In fact, determining the OC is more difficult

## Chapter 1

---

in the presence of state and/or control constraints; therefore numerical methods must be employed.

During our study of computational OC theory, we have surveyed many practical computing methods which have been developed for solving OC problems since the last century. It can be acknowledged that the majority of the presented methods in the literature successfully solve the unconstrained problems, but the presence of the state/control variables inequality constraints usually lead to both analytical and computational difficulties. Most of the numerical methods presented for solving OC problems generally fall into three classes: Numerical DP, direct optimization methods (DOMs) and indirect optimization methods (IOMs). The latter two classes of methods were originally labeled by von Stryk and Bulirsch in 1992. The recent computational advances in the solution of OC problems evident in numerous research papers and textbooks favor DOMs over the other two approaches. Indeed, despite the advances in computational power and digital computers, numerical DP and IOMs continue to suffer from fundamental problems which limit their applications for solving general nonlinear OC problems. In particular, numerical DP attempts to solve OC problems by numerical backward induction suited for the treatment of integer variables, but suffers in general from the numerical difficulties related to the “curse of dimensionality,” which renders the HJB equation “impossible to solve in most cases of interest” (Polak, 1973). Therefore it is not the method of choice for generic large-scale OC problems with underlying nonlinear dynamical system of equations. In fact, this class of computational methods had no significant echo in the community of computational OC theory in the recent decades, and did not attract much attention to the extent that many authors like Elnagar and Kazemi (1998a); Fahroo and Ross (2000); Gong et al. (2006a); Sager et al. (2009); von Stryk and Bulirsch (1992) have overlooked its existence, and considered DOMs and IOMs as the two most general numerical approaches for solving OC problems.

The IOMs are known as “first optimize, then discretize approaches,” since the optimality conditions are found first before the application of the numerical techniques. These methods were applied in the early years of solving OC problems numerically. In these methods, CV and the MP are applied to determine the first-order necessary conditions for an optimal solution (Bryson and Ho, 1975; Kirk, 2004). In this manner, the OC problem is transformed into a two-point boundary-value problem (TPBVP). The IOMs then endeavor to find an approximate solution to the generally nonlinear TPBVP by iterating on it to seek its satisfaction. The term “indirect” is coined with this class of methods since the OC is determined by solving the auxiliary TPBVP rather than directly focusing on the original problem (Subchan and Zbikowski, 2009). Although IOMs may produce a highly accurate approximation to the OC, they are generally impractical. In fact, finding numerical solutions to the nonlinear TPBVP is ex-

## Chapter 1

---

tremely difficult, and the IOMs suffer from numerical difficulties associated with the stiffness of the augmented system of differential equations. Moreover, since the solution of the TPBVP entails the integration of the state differential equations forward in time, followed by the integration of the corresponding costate differential equations backward in time, the solution of the augmented system is carried out in opposite directions; thus it is very hard to implement an adaptive integration scheme, since it is almost impossible to ensure that the state and costate discretization points sets coincide. Hence the solution scheme must invoke an appropriate interpolation method, since the solution of the costate system is dependent on the solution of the state system. The matter which compromises the accuracy of the numerical scheme (Liu, 2011). Furthermore, more difficulties are conspicuous particularly for problems with interior point constraints (Ghosh et al., 2011). From another viewpoint, IOMs require the derivation of the complicated first-order necessary optimality conditions, which include the adjoint equations, the control equations, and all of the transversality conditions. These necessary conditions of optimality are difficult to formulate for problems of even simple to moderate complexity (Darby, 2011), and are quite daunting for intricate OC problems. Moreover, modifying a component of the OC problem such as adding or removing a constraint entails the reformulation of the necessary conditions. Another difficulty arises for problems whose solutions have active path constraints. In these cases, a priori knowledge of the switching structure of the path constraints must be known (Darby, 2011; Garg, 2011). In fact, the priori estimate of the constrained-arc sequence may be quite difficult to determine, which makes it extremely difficult to impose the correct junction conditions and define the arc boundaries (Betts, 2001). Betts (1998) pointed out that in the cases where inequality constraints are imposed, additional jump conditions imposed at the junction points must be satisfied extending the TPBVP into a multi-point boundary value problem (BVP). Moreover, one must model the IOMs in multiple-phases, since the optimality conditions are different on the constrained and the unconstrained arcs (Betts, 2009). Furthermore, these methods are known to be quite unstable (hyper-sensitive), since they suffer from their inherent small radii of convergence as the necessary conditions of optimality usually lead to a stiff TPBVP, which must be solved to obtain the approximate solutions (Rao, 2003). To make things worse, a serious drawback is that the user must guess values for the adjoint variables (unreasonably good initial guesses in many cases), which is very non-intuitive as they are not physical quantities (Betts, 2009; Darby, 2011; Gao and Li, 2010; Lin et al., 2004; Padhi et al., 2006; Yang, 2007), and the MP gives no information on the initial value of the costates (Liu, 2011). Hence providing such a guess is difficult. Even with a reasonable guess for the adjoint variables, the numerical solution of the adjoint equations can be very ill-conditioned; thereby the Hamiltonian generates a numerically sensitive BVP that may produce wild

## Chapter 1

---

trajectories which exceed the numerical range of the computer (Bryson and Ho, 1975; Ross and Fahroo, 2003). Besides all of these concerns, a solution of an IOM may not be a local minimum or a global minimum, since the MP provides the necessary conditions of optimality, which are not sufficient in general for nonlinear OC problems. This indicates that once all the extrema of a given OC problem are found, one may search among them to locate the OC law.

The DOM as a third generic approach views the continuous-time optimal control problem (CTOCP) as an infinite-dimensional optimization problem, such that the optimization solver searches for the control function which optimizes the objective functional while satisfying certain equality/inequality constraints. Since the optimization routines do not operate on infinite-dimensional spaces, the control and the state variables are approximated/parameterized first before the outset of the optimization process. Therefore, the terms “first discretize, then optimize approach,” or “all-at-once approach” are usually coined with a DOM. The transcription of the infinite-dimensional CTOCP into a finite-dimensional nonlinear programming (NLP) problem is carried out through discretization of the original problem in time and performing some parameterization of the control and/or state vectors. The resulting static NLP problem is then solved using some well-developed optimization methods and software. The convergence to a solution of the CTOCP is usually accomplished by taking a finer mesh in the discretization scheme, and the optimal values of the states and controls are typically obtained at the discrete time points.

### 1.1.4 Advantages of DOMs

The verification of optimality in the DOMs may not be an easy task. Nonetheless, there are a number of advantages associated with the implementation of DOMs which render them more practical than IOMs. Firstly, an important benefit of recasting the problem as a NLP problem is that it eliminates the requirement of solving a stiff TPBVP, and finding a closed-form expression for the necessary and sufficient conditions of optimality, which are calculated off-line in the MP and DP methods, respectively. Therefore DOMs can be quickly used to solve many practical trajectory optimization problems in a short time (Fahroo and Ross, 2002). Secondly, it is easy to represent the state and the control dependent constraints; thus the OC problem can be modified easily. Moreover, DOMs tend to have better convergence properties with no requirements to guess the costate variables. In contrast, the IOMs exhibit small radius of convergence requiring sufficiently close initial guesses (Fahroo and Ross, 2002). Additionally, DOMs can be formulated using a single phase, while IOMs are modeled using multiple phases, whereas the optimality conditions are different on the constrained and the unconstrained arcs (Betts, 2009). These features together with the simplicity of the discretiza-

## Chapter 1

---

tion procedure, the high accuracy, and the fast convergence of the solutions of the discretized OC problem to the solution of the underlying infinite-dimensional OC problem have made the DOMs the ideal methods of choice nowadays (Gong et al., 2006a, 2008; Hesthaven et al., 2007), and well-suited for solving OC problems; cf. (Benson, 2004; Benson et al., 2006; Betts, 1998, 2009; Chen et al., 2011; El-Gindy et al., 1995; El-Hawary et al., 2003; Elnagar et al., 1995; Elnagar and Zafiris, 2005; Elnagar, 1997; Elnagar and Kazemi, 1995, 1998a,b; Elnagar and Razzaghi, 1997; Fahroo and Ross, 2002, 2008; Garg et al., 2011a,b, 2010; Gong et al., 2006a; Hargraves and Paris, 1987; Hull, 1997; Huntington, 2007; Jaddu, 2002; Kang and Bedrossian, 2007; Kang et al., 2007, 2008; Razzaghi and Elnagar, 1993, 1994; Stryk, 1993; Vlassenbroeck and Dooren, 1988; Williams, 2004), and the references therein.

### 1.1.5 Elements of DOMs and the Shooting Methods

In a DOM, three elements need to be considered, viz. (Kaya, 2010) (i) the choice of the discretization scheme, (ii) the convergence of the optimization technique employed, and (iii) the convergence to a solution of the OC problem as one takes a finer mesh in the discretization scheme. Fortunately, there are many DOMs in the literature which can be applied for the solution of the OC problems. The direct single-shooting methods and the direct multiple-shooting methods are two of the earliest DOMs for solving OC problems (Betts, 2009; Bock and Plitt, 1984). In both DOMs, the control variables are parameterized using some functional form, and the dynamical system is solved for the state variables using some time-marching scheme, for instance. In fact, the shooting methods have been originally used extensively for the solution of TPBVPs (Betts, 2009; Burden and Faires, 2000; Cheney and Kincaid, 2012; Süli and Mayers, 2003). In the single-shooting method, an initial value problem (IVP) corresponding to the TPBVP is solved instead. The initial data of the IVP is adjusted in a certain way so that the solution of the IVP finally satisfies the TPBVP. In fact, a direct single-shooting method may be useful for OC problems approximated using relatively small number of variables; however, the success of the direct single-shooting methods can be degraded significantly for increasing numbers of the variables (Pytlak, 1999). Moreover, due to the instabilities of the IVP solution over longer time intervals, single-shooting methods may “blow up” before the IVP can be completely integrated. In fact, this can appear even with extremely accurate guesses for the initial values. Hence direct single-shooting methods are rather impractical for solving a wide variety of OC problems. In an attempt to resolve these difficulties, multiple-shooting methods for the solution of TPBVPs were proposed by Morrison et al. (1962), while direct multiple-shooting methods were introduced by Bock and Plitt (1984) for the solution of OC problems. The key idea of these meth-

## Chapter 1

---

ods is to break up the TPBVP into a system of coupled TPBVPs over smaller time subintervals whose endpoints are called nodes. In this manner, the single-shooting method can be applied over each subinterval through the use of some state continuity conditions (matching conditions) enforced between consecutive subintervals rather than just at the boundary points of the solution interval. The divisions of the TPBVP solution domain into several subintervals decrease the sensitivity of the multiple-shooting methods to the initial guess. Therefore the direct multiple-shooting methods are more stable and offer improvements over the standard direct single-shooting methods. Yet, both direct shooting methods are computationally expensive due to the numerical integration operations and the requirement of a priori knowledge of the switching structure of inactive and active path constraints. The reader may consult (Betts, 1998, 2009; Bock and Plitt, 1984; Pytlak, 1999; von Stryk and Bulirsch, 1992) for an overview of the direct shooting methods.

### 1.1.6 Typical Discretization Methods for Dynamical Systems

The discretization of the dynamics in a DOM can be carried out using various numerical methods. The common numerical schemes are the finite difference methods, such as Euler's method and Runge-Kutta methods, finite element methods, piecewise-continuous polynomials such as linear splines, cubic splines, or B-spline methods, interpolating scaling functions, wavelets methods such as Walsh-wavelets and Haar wavelets approximations, block pulse function methods, etc.; cf. (Ait and Mackenroth, 1989; Becker and Vexler, 2007; Bonnans and Laurent-Varin, 2006; Chen and Lu, 2010; Dontchev and Hager, 1997; Dontchev et al., 2000; Foroozandeh and Shamsi, 2012; Glabisz, 2004; Hargraves and Paris, 1987; Hsiao, 1997; Hwang et al., 1986; Kadalbajoo and Yadaw, 2008; Kaya and Martínez, 2007; Kiparissides and Georgiou, 1987; Lang and Xu, 2012; Liu et al., 2004; Liu and Yan, 2001; Pytlak, 1998; Schwartz and Polak, 1996; Stryk, 1993; Xing et al., 2010). However, a common feature in these approximation methods is that they usually experience an explosion in the number of variables if high orders of accuracy are sought, except for very specialized cases where the control is of a bang-bang control type (Kaya, 2010). This is due to the finite-order convergence rates associated with these methods (Weideman and Reddy, 2000).

## 1.2 Spectral Methods

Spectral methods, amongst the available discretization methods in the literature, offer useful alternatives to the aforementioned methods for solving differential

## Chapter 1

---

equations, eigenvalue problems, optimization problems, OC problems, and in many other applications; cf. (Benson et al., 2006; Boyd, 2001; Elgindy, 2009; Elgindy and Hedar, 2008; Elnagar et al., 1995; Fahroo and Ross, 2002; Gong et al., 2006a; Kang and Bedrossian, 2007; Mason and Handscomb, 2003; Quarteroni and Valli, 1994; Ross and Fahroo, 2003). The number of advantages of spectral methods which can be found in numerous textbooks, monographs, and research papers are extensive, yet we shall mention some few benefits of applying spectral methods which are quite useful for the work conducted in this dissertation: (i) One conspicuous advantage of spectral methods is that they are memory minimizing as they achieve high precision accurate results using substantially fewer grid points than required by typical finite difference schemes and other methods (Zang et al., 1982). (ii) The previous elegant task is accomplished while providing Eulerian-like simplicity (Gong et al., 2007); therefore the spectral computations can be considerably more effective (Barranco and Marcus, 2006). (iii) Another significant advantage of spectral methods is that the boundary conditions imposed on the spectral approximations are normally the same as those imposed on the differential equation. On the other hand, finite-difference methods of higher order than the differential equation require additional “boundary conditions” (Gottlieb and Orszag, 1977). (iv) Besides these desirable features, spectral methods produce global solutions, rapid convergence, and most of all, they are highly accurate for problems exhibiting smooth solutions to the extent that they are often used in cases when “nearly exact numerical solutions are sought” (Cushman-Roisin and Beckers, 2011; Gardner et al., 1989). Hence spectral methods are able to approximate the cost function, the dynamical system, the states and the controls constraints functions very precisely when the solutions are smooth, which is absolutely vital for the successful approximation of OC problems. These advantages place spectral methods at the front of the available numerical methods for solving CTOCPs exhibiting sufficiently differentiable solutions. The class of discontinuous/nonsmooth OC problems can be treated by spectral methods as well through some special techniques, which are highlighted in Chapters 6 and 8.

### 1.2.1 Orthogonal Collocation/Pseudospectral (PS) Methods

Among the available spectral methods in the literature, orthogonal collocation/PS methods, which use orthogonal global polynomials, have emerged as important and popular computational methods for the numerical solution of OC problems in the last two decades, and their advantageous properties are highlighted by many authors; cf. (Benson, 2004; Benson et al., 2006; Elnagar et al., 1995; Elnagar, 1997; Elnagar and Razzaghi, 1997; Fahroo and Ross, 2008; Garg et al.,

## Chapter 1

---

2011b, 2010; Huntington, 2007; Rao et al., 2010; Williams, 2004). They have been universally applied in many applications because of their greater simplicity and the computational efficiency compared to the other spectral methods, viz. Galerkin and Lanczos tau methods (Gottlieb and Orszag, 1977). In fact, the popularity of the orthogonal collocation methods is largely due to their ability to offer an exponential convergence rate for the approximation of analytic functions while providing an Eulerian-like simplicity. Consequently, these methods can generate significantly smaller-scale optimization problems compared to traditional discretization methods.

### 1.2.2 Direct Orthogonal Collocation Methods (DOCMs) and Direct PS Methods

The application of the orthogonal collocation/PS methods for the solution of OC problems gives rise to the efficient class of methods known as the DOCMs, direct PS methods, or direct transcription methods. In the DOCMs, both the state and the control are parameterized using a set of global orthogonal trial (basis) functions. The cost function is approximated in an algebraic form, and the dynamical system given in the form of differential-algebraic constraints is enforced at a finite number of points called the collocation points. The dynamics is imposed at the collocation points by ways of numerical differentiation/integration operators usually referred to by the differentiation/integration matrices. In this manner, the OC problem is transformed into a constrained NLP problem, which can be solved using the powerful optimization methods. Convergence to the exact solution is achieved by increasing the degree of the polynomial approximation of the state and the control variables. It is worthy to mention that in the recent years, DOCMs have emerged as demonstrable candidates for real-time computation. Before that, the computational methods were widely considered as being too slow for real-time applications of highly nonlinear problems (Kang et al., 2008). Moreover, DOCM/PS methods are very useful in estimating highly accurate costate/adjoint variables approximations of the indirect problem. In fact, this intriguing result has been investigated by Benson et al. (2006); Fahroo and Ross (2001); Garg et al. (2011b) by determining the relationship between the costates associated with the TPBVP of the indirect problem and the Lagrange (Karush-Kuhn-Tucker) multipliers of the direct problem. Recently, Gong et al. (2008); Kang et al. (2008) have also provided the results on the existence and the convergence of the solution of the discretized OC problem to that of the original OC problem using Legendre-Gauss-Lobatto (LGL) PS methods.

## Chapter 1

---

### 1.2.3 DOCMs Versus Shooting Methods and IOMs

DOCMs present more computational efficiency and robustness over the shooting methods and IOMs for solving OC problems. In fact, in a DOCM, one usually establishes a well-behaved numerical scheme by parameterizing both the state and the control variables, while the shooting methods are dubbed as control parameterization approaches, since the control variables are only parameterized. The state variables are then obtained by solving the dynamical system using the available ordinary differential equation (ODE) solvers. This numerical scheme is usually intensive computationally and sensitive to numerical errors (Yen and Nagurka, 1992). Moreover, in contrast to IOMs and direct shooting methods, DOCMs do not require a priori knowledge of the active and inactive arcs for problems with inequality path constraints (Darby, 2011; Garg, 2011), and the user does not have to be concerned with the adjoint variables or the switching structures to determine the OC (von Stryk and Bulirsch, 1992). Furthermore, DOCMs show much bigger convergence radii than either the indirect methods or direct shooting methods as they are much less sensitive to the initial guesses. The memory minimizing feature of the orthogonal collocation approximations is highly relevant in a direct optimization scheme as it results in a finite-dimensional NLP with considerably lower dimension compared to other competitive methods in the literature. Moreover, to quote Polak (2011, pg. 251): ‘current nonlinear optimization algorithms ... usually solve the problems discretized using collocation techniques much more rapidly than when applied to “classical” discretizations.’ One of the significant applications of spectral collocation methods which has received wide publicity recently was in generating real time trajectories for a NASA spacecraft maneuver (Kang and Bedrossian, 2007).

### 1.2.4 DOCMs Versus Standard Direct Local Collocation Methods (DLCMs)

DOCMs are different from the conventional DLCMs in the sense that global orthogonal basis functions are employed, while the latter use fixed low-degree approximations after dividing the solution interval into many subintervals. In fact, DLCMs are typically developed as local methods, where second-order polynomials and piecewise-continuous polynomials such as linear or cubic splines over each time segment are some of the interpolating polynomials used (Hargraves and Paris, 1987; Stryk, 1993; Tsang et al., 1975). Yet the class of Runge-Kutta methods was most often used (Betts, 1998, 2009; Hager, 2000). An interpolation scheme is used to obtain the time histories of both the control and the state variables. Moreover, the convergence of the numerical discretization is achieved by increasing the number of subintervals (Betts, 2009). Although standard DLCMs

## Chapter 1

---

result in sparse NLP problems, the convergence is at a polynomial rate. Therefore an excessively large number of subintervals may be required to accurately approximate the solution. In contrast, the convergence to the exact solution in a DOCM is achieved by increasing the degree of the polynomial approximation usually in a single time interval. Once the approximate state and control variables are approximated at a given collocation nodes set, the time histories of both the state and the control variables can be easily determined directly without invoking any interpolation methods. Moreover, for problems whose solutions are smooth and well-behaved, DOCMs converge at an exponential rate (Benson, 2005; Garg et al., 2010; Gong et al., 2006a). Consequently, they usually approximate the solutions of the OC problem using a few number of NLP variables.

### 1.2.5 Orthogonal Functions

Orthogonal functions have been used abundantly in the literature as the expansion basis for the solution of OC problems. The main feature of the methods based on orthogonal series expansions is that they reduce the differential equations constraints of the OC problem into algebraic or transcendental equations in terms of the expansion coefficients of the unknown state/control functions, often through the operational matrices of differentiation/integration. This attractive property found much appealing by many authors, since it greatly simplifies the OC problem, and ultimately renders the solution within an easy reach of advanced linear/nonlinear algebra and optimization tools. Typical examples of the orthogonal functions which have been applied in the literature of computational OC theory are the Walsh functions (Aoki, 1960; Chen and Hsiao, 1975; Palanisamy and Prasada, 1983), the block pulse functions (Hsu and Cheng, 1981; Hwang et al., 1986; Rao and Rao, 1979), Laguerre polynomials (Horng and Ho, 1985; Hwang and Shih, 1983), Chebyshev polynomials (El-Gindy et al., 1995; Elnagar and Kazemi, 1998a; Liu and Shih, 1983; Paraskevopoulos, 1983), Hermite polynomials (Coverstone-Carroll and Wilkey, 1995; Kekkeris and Paraskevopoulos, 1988), Fourier series (Endow, 1989; Nagurka and Yen, 1990; Razzaghi, 1990; Yang and Chen, 1994), Legendre polynomials (Elnagar et al., 1995; Razzaghi and Elnagar, 1993; Ross and Fahroo, 2003; Shyu and Hwang, 1988; Yan et al., 2001), shifted Legendre polynomials (Hwang and Chen, 1985; Shih and Kung, 1986), shifted Chebyshev polynomials (Chou and Horng, 1985a,b; Razzaghi and Razzaghi, 1990), Gegenbauer (ultraspherical) polynomials (El-Hawary et al., 2003; Lee and Tsay, 1989; Tsay and Lee, 1987), and Jacobi polynomials (Lee and Tsay, 1989; Tsay and Lee, 1987; Williams, 2004).

## Chapter 1

---

### 1.2.6 Jacobi Polynomials

For smooth periodic problems, Fourier spectral methods, which employ the periodic basis  $e^{inx}$ , have been demonstrated to perform well and achieve very precise approximations. However, the exponential accuracy of these methods is only guaranteed when the solution and all of its derivatives are periodic functions. The lack of such globally periodicity conditions significantly reduce the convergence rate of the Fourier series. In fact, it is well-known that the truncated Fourier series of a nonperiodic function having a discontinuous periodic extension at the boundaries, converges very slowly, like  $O(1/N)$ , inside the region, and produce  $O(1)$  spurious oscillations near the boundaries— a phenomenon known as the Gibbs phenomenon (Vozovoi et al., 1996). Since satisfying the periodicity constraints of the solution function and its derivatives are rather very tight, and not valid in many applications, it is naturally intuitive to consider spectral expansions based on non-periodic bases, such as the polynomial bases. The convergence properties of the discrete solutions of OC problems discernible in many research works in the literature of computational OC theory clearly favor the Jacobi class of polynomials over the other classes of orthogonal functions. This family of polynomials has been extensively studied and exhibits very nice convergence properties. In fact, for nonperiodic problems, Jacobi polynomials have been the most successful orthogonal basis polynomials by far (Fornberg, 1996), and the expansion in the Jacobi polynomials is accurate independent of the specific boundary conditions of the solution function (Hesthaven et al., 2007). The Jacobi family of polynomials  $P_n^{(\alpha,\beta)}(x)$  of degree  $n$  and associated with the real parameters  $\alpha; \beta$  include the Gegenbauer polynomial  $C_n^{(\lambda)}(x)$  of degree  $n$  and associated with the real parameter  $\lambda$  for values of  $\alpha = \beta = \lambda$ . The latter include the Chebyshev polynomial of the first kind  $T_n(x)$ , and Legendre polynomial  $L_n(x)$  for  $\lambda = 0; 0.5$ , respectively (Boyd, 2006).

### 1.3 Motivation of the Present Work

Although DOCMs and direct PS methods have been applied successfully in many OC applications, these methods typically employ the square and dense spectral differentiation matrices (SDMs) to approximate the derivatives arising in the OC problem. Therefore, there are two clear limitations associated with these methods: (i) SDMs are known to be ill-conditioned (Funaro, 1987), and suffer from many drawbacks; cf. Figure 1.1; therefore, the approximate solutions obtained by these methods may be sensitive to the round-off errors encountered during the discretization procedure. (ii) To determine approximations of higher-orders, one usually has to increase the number of collocation points in a direct PS method,

## Chapter 1

---

which in turn increases the number of constraints and the dimensionality of the resulting NLP problem; cf. Figure 1.2. Also increasing the number of collocation points in a DOCM increases the number of constraints of the reduced NLP problem. Eventually, the increase in the size of the SDM leads to larger NLP problems, which may be computationally expensive to solve and time-consuming.

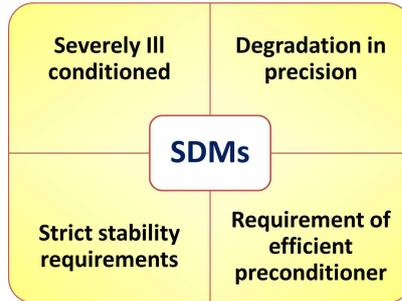


Figure 1.1: SDMs are known to be severely ill-conditioned, and their implementation causes degradation of the observed precision. Moreover, it has been shown that the time step restrictions can be more severe than those predicted by the standard stability theory (Trefethen, 1988; Trefethen and Trummer, 1987). For higher-order SDMs, the development of efficient preconditioners is extremely crucial (Elbarbary, 2006; Hesthaven, 1998).

The research work presented in this dissertation is prompted by the need to generate robust solutions to complex OC problems quickly and accurately, with limited memory storage requirements, and without increasing the size of the resulting NLP problem. We shall demonstrate later in this dissertation that a well-behaved numerical scheme enjoying these useful features can be established by recasting the dynamics into its integral formulation and working under complete integration framework. Moreover, the choice of the orthogonal basis polynomials employed in a DOCM is another crucial element in the path of planting a strong and practical OC solver. In fact, the choice of the optimal orthogonal basis polynomials among the Jacobi polynomials employed for solving CTOCPs remains the subject of intense debate. Chebyshev and Legendre polynomials have been frequently applied as the bases functions for DOCMs since they were originally proposed by Vlassenbroeck and Dooren in 1988 and Elnagar et al. in 1995, respectively. Part of the reason is due to their fast convergence behaviors exhibited in many applications. Moreover, these two bases polynomials remain well-conditioned for expansions using hundreds or thousands of the spectral expansion terms. The work established in this dissertation confirms these arguments if Chebyshev and Legendre polynomial expansions were to be compared with other members of the Jacobi family of polynomials for large numbers of the

## Chapter 1

---

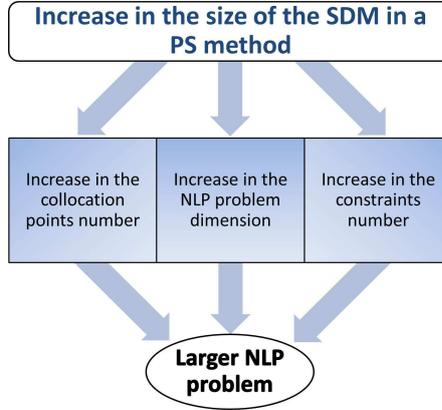


Figure 1.2: Any increase in the size of the square SDM in a direct PS method requires the same increase in the number of collocation points, which leads to an increase in the dimension of the NLP problem, and the number of constraints. Eventually, these elements accumulate to yield a larger NLP problem.

spectral expansion terms. But “*what about short/medium-range expansions of the Chebyshev and Legendre polynomials compared to the Gegenbauer polynomials or other members of the Jacobi polynomials?*” This important question, to the best of our knowledge, has not been highlighted thoroughly in the literature of spectral methods. This question is quite substantial in many applications including OC theory, since a few numbers of parameters are always highly desirable for parameterizing the controls and/or the states in a DOCM, and to establish a small-scale NLP problem accurately representing the original OC problem. Moreover, the small number of spectral expansion terms used in the approximation has been always a strong measure for the quality of the discretization and the efficiency of DOCMs/PS methods; cf. (El-Gindy et al., 1995; Elnagar and Kazemi, 1995; Garg et al., 2010; Huntington, 2007; Ma et al., 2011; Razzaghi and Elnagar, 1993; Vlassenbroeck and Dooren, 1988), etc.

In this dissertation, we shall investigate the application of the more general family of polynomials, the Gegenbauer polynomials. These polynomials have received much attention in the literature, especially in applied mathematics for their fundamental properties and versatile applications (Keiner, 2009). In fact, there are a number of reasonable arguments drawn from the rich literature of spectral methods which have motivated us to apply the more general Gegenbauer family of polynomials, at least for short-range expansions of the spectral basis polynomials employed in the approximations. In the following, we shall briefly mention a few reasons considerably important for our focus on the Gegenbauer polynomials as

## Chapter 1

---

an apt choice of orthogonal basis polynomials used for the solution of CTOCPs: (i) Light (1978) computed the norms of a number of Gegenbauer projection operators finding that they all increased monotonically with the Gegenbauer parameter  $\lambda$  so that the Chebyshev projection cannot be minimal (Mason and Handscomb, 2003). In particular, the reported results show that the norm of the Chebyshev projection is not the smallest for Chebyshev series expansions truncated after  $n$  terms in the range  $1 < n < 10$ . (ii) The Gegenbauer expansion method presented by Doha (1990) to numerically approximate the solution of BVPs of linear partial differential equations (PDEs) in one dimension shows that more accurate results are obtained by taking  $\lambda$  small and negative. Moreover, the precision of the Gegenbauer polynomials approximations exceeded those obtained by the Chebyshev and Legendre polynomials. (iii) Expansions in Chebyshev polynomials are better suited to the solution of hydrodynamic stability problems than expansions in other sets of orthogonal functions (Orszag, 1971). On the other hand, in the resolution of thin boundary layer applications, Legendre polynomials expansion gives an exceedingly good representation of functions that undergo rapid changes in narrow boundary layers. Hence, it is convenient to apply a unified approach using the Gegenbauer polynomials, which include Chebyshev and Legendre polynomials as special cases, rather than applying a particular polynomial for various approximation problems. Moreover, the theoretical and experimental results derived in a Gegenbauer solution method apply directly to Chebyshev and Legendre polynomial approximation methods as special cases. (iv) The numerical treatment of definite integrations applied by El-Hawary et al. (2000) using Gegenbauer integration matrices (GIMs) followed by their application for the numerical solution of OC problems (El-Hawary et al., 2003) revealed many useful advantages over the standard choices of Chebyshev and Legendre polynomial expansion methods. (v) Finally, the recent work of Imani et al. (2011) on the solution of nonlinear BVPs shows that the Jacobi polynomials generally produce better results compared to Chebyshev and Legendre polynomials for various values of the parameters  $\alpha; \beta$ .

It is noteworthy to mention that the work of Williams (2004) showed that the Jacobi polynomials can be very efficient in the solution of OC problems. Nonetheless, his numerical example of the orbit transfer problem showed that a wide gap in the calculation time occurs as the parameters of the Jacobi polynomial change, and it is recommended to opt for the values of  $\alpha; \beta$  which give the most efficient computation time for a specific application. Williams though did not provide any clue or perhaps a rule of thumb for choosing optimal (or loosely speaking, appropriate) choices of the Jacobi parameters to establish an efficient solution method. On the other hand, the El-Hawary et al. (2003) work in the solution of OC problems using Gegenbauer polynomials seems to offer a suitable method for optimally choosing the Gegenbauer parameter  $\lambda$  in a certain application. The

## Chapter 1

---

latter article as well as the absence of a valid numerical method for determining suitable Jacobi parameters have motivated us further to pursue the solution of CTOCPs using Gegenbauer expansion series.

### 1.3.1 Collocation Points

The choice of the discretization/collocation points plays an important role in the solution of OC problems as well. In fact, it can be easily demonstrated that an arbitrary choice of the collocation nodes can deliver very poor approximations. This may cause severe problems such as the Runge phenomenon<sup>1</sup>, which is known as the divergence of the interpolants constructed using equispaced-interpolation-points at the endpoints of the interpolation domain. Therefore, the impact of the collocation points choice on the performance and the efficiency of the DOCMs is parallel to the effect of the orthogonal basis polynomials used in the approximation. Several results in approximation theory have shown that different collocation points sets of the Gauss type yield superior interpolation approximations of functions to the ones obtained from equidistant points. These points sets have the distribution property of clustering around the endpoints of the interval, which results in the avoidance of the Runge phenomenon (Trefethen, 2000). In fact, it can be shown that the interpolation errors decrease exponentially for interpolation based on Gauss points. The derivatives/integrations of the interpolating polynomials at these points sets can be obtained exactly through differentiation/integration matrices. The choice of the orthogonal basis functions and the Gauss collocation points separate the orthogonal collocation/PS methods from the other collocation methods in the literature.

The most well-developed DOCMs classified according to the choice of the collocation nodes are the Gauss, Gauss-Radau, and Gauss-Lobatto DOCMs; cf. (Benson, 2004; Benson et al., 2006; Elnagar et al., 1995; Fahroo and Ross, 2001; Garg et al., 2011a,b, 2010; Gong et al., 2006a, 2009; Rao et al., 2010; Williams, 2006). The collocation points sets in these methods are the Gauss, Gauss-Radau, and the Gauss-Lobatto points, respectively. These points sets are defined on the interval  $[-1, 1]$ , and they chiefly differ in how the end points are incorporated. The Gauss points set does not include both endpoints. The Gauss-Radau points set include one endpoint only, while the Gauss-Lobatto points set include both endpoints. The latter appears to be the most intuitive points set to be incorporated, since OC problems generally have boundary conditions at both the initial and terminal times. However, many recent research articles show that they may not be the most appropriate. Garg et al. (2010) showed that the differentiation matrices of the Legendre-Gauss (LG) and Legendre-Gauss-Radau (LGR) schemes, for

---

<sup>1</sup>This phenomenon was originally discovered by Carl David Tolmé Runge (1856–1927).

## Chapter 1

---

instance, are rectangular and full rank, whereas the LGL differentiation matrix is square and singular. Consequently the LG and LGR methods can be written equivalently in either differential or implicit integral form, while the LGL method does not have an equivalent implicit integral form. Moreover, the LG and LGR transformed adjoint systems are full rank while the LGL transformed adjoint system is rank-deficient. Consequently, the LG and LGR costate approximations converge exponentially while the LGL costate is potentially non-convergent. Also it was shown that the costate estimate using LGL points set tends to be noisy due to the oscillations about the exact solution, and these oscillations have the same behavior as the null space of the LGL approximation (Darby, 2011). Benson (2004); Garg et al. (2010) showed further that the LG and LGR methods produce highly accurate discrete approximations to the solutions of the OC problem.

### 1.3.2 Choice of the NLP Solver

The above arguments sustain the application of the Gegenbauer orthogonal collocation method together with the Gauss/Gauss-Radau collocation nodes as a suitable numerical scheme competent to efficiently discretize CTOCPs. Since the solution of OC problems using DOCMs is established through two key stages: The numerical discretization scheme and the optimization method employed for solving the resulting constrained NLP problem, it is substantial to consider an efficient optimization method for solving the reduced optimization problem. In this dissertation, we shall apply the interior-point method provided with MATLAB software. This popular optimization method reduces the original inequality-constrained problem into a sequence of equality constrained problems, which are easier to solve. Moreover, the interior-point method is well-known for its simplicity and its ability to generally solve a large optimization problem in a small number of iterations (Li and Santosa, 1996). Furthermore, the interior-point method converges in polynomial time, and provides scaling to handle the numerical ill-conditioning arising during the numerical computations (Ahmadi and Green, 2009; Singh et al., 2008).

### 1.3.3 Significance of the Operational Matrix of Integration

In addition to the spectral orthogonal basis polynomials used in the approximation, the collocation points type of sets used for the discretization of the OC problem, and the optimization solver employed for the solution of the reduced NLP problem, one must acknowledge that the precision and the stability of the numerical differentiation/integration operators applied for approximating the cost function, the dynamical system, and the state/control constraints are extremely

## Chapter 1

---

substantial ingredients for the establishment of an efficient DOCM. In fact, these two properties, namely the precision and the stability, are crucial factors in determining the dimension of the resulting NLP problem. For instance, a poor spectral differentiation/integration operator may require many spectral expansion terms and collocation points to accurately represent the differentiation/integration operators involved in the OC problem. This matter poses a problem when the memory storage is limited, since the spectral solution of a large-dimensional NLP problem often requires the storage of large and dense matrices. In contrast, precise and well-conditioned spectral operators can deliver a significantly small-scale optimization problem. This feature is very important as it facilitates the task of the optimization solver, and significantly reduces the required computational time. Due to the lack of stability in the SDMs, especially those of high-orders as evident in many research works of this area; cf. (Boyd, 2001; Elbarbary, 2006; Greengard, 1991; Tang and Trummer, 1996; Trefethen, 1996), we shall recast the OC problem into its integral form, and consider the application of the Gegenbauer integration matrices for approximating the integration operators involved in the OC problem.

It is noteworthy to mention that despite the fact that spectral methods are represented by dense/full matrices, ‘explicit time-stepping algorithms can be implemented nearly as efficiently for them as for finite difference methods on a comparable grid’ (Zang et al., 1982).

### 1.4 Framework

In this dissertation, we present a fast and efficient Gegenbauer transcription method (GTM) for solving CTOCPs based on Gegenbauer-Gauss (GG) collocation. The proposed method is a DOCM which parameterizes the state and the control variables using global Gegenbauer polynomial expansions, and solves the OC problem directly for the states and controls. For problems with various orders time derivatives of the states arising in the cost function, dynamics, or constraints, the GTM solves the CTOCP directly for the control  $u(t)$  and the highest-order time derivative  $x^{(N)}(t)$ ,  $N \in \mathbb{Z}^+$ . The state vector and its derivatives up to the  $(N - 1)$ th-order derivative can then be stably recovered by successive integration. To take full advantage of the useful Gegenbauer integration matrices operators, we shall transform the dynamical system of differential equations into its integral formulation. In this manner, a Bolza CTOCP is transformed into an integral Bolza CTOCP. The transformed dynamical system can be imposed by discretization at the GG points using an optimal GIM which exploits the strengths of the Chebyshev, Legendre, and Gegenbauer polynomials, and possesses many advantages; cf. Figure 1.3. The novel optimal GIM introduced in this dissertation is

## Chapter 1

---

constructed through interpolations at some optimal set of GG points adjoining the solution points so that the Gegenbauer quadrature error is minimized in a certain optimality measure. This technique entails the calculation of some locally optimal Gegenbauer parameter values  $\{\alpha_i\}_{i=0}^{N_\alpha}$ ,  $N_\alpha \in \mathbb{Z}^+$ , rather than choosing any arbitrary  $\alpha$  value. In this manner, we can improve the quality of the discrete Gegenbauer approximations, and significantly reduce the dimension of the resulting NLP problem. The integral cost function can be discretized also by the optimal GIM, which provides highly accurate results for approximating definite integrals. Hence the present method unifies the process of the discretization of the differential equations and the integral cost function; cf. Figure 1.4. We restrict ourselves to CTOCPs whose dynamics are described by ODEs, and/or provided with state/control constraints. CTOCPs governed by integro-differential equations are solved similarly by recasting the dynamics into its integral formulation, while CTOCPs with integral equation constraints are solved straightforwardly.

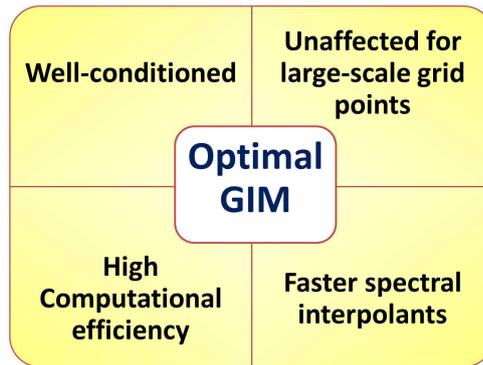


Figure 1.3: The optimal GIM captures the most suitable properties of the Chebyshev, Legendre, and Gegenbauer polynomials required for a given problem. It is a well-conditioned operator, and its well-conditioning is essentially unaffected for increasing number of grid points. The use of integration for constructing the spectral approximations improves the rate of convergence of the spectral interpolants, and allows the multiple boundary conditions to be incorporated more efficiently.

### 1.4.1 Advantages of the GTM

The GTM derived in this dissertation has many advantages over other DOCMs/PS methods for solving OC problems. (i) The presented GTM can be easily programmed, consistently preserves the type of the constraints, and encompasses a wider range of OC problems than the usual DOMs. The GTM produces a NLP

## Chapter 1

---

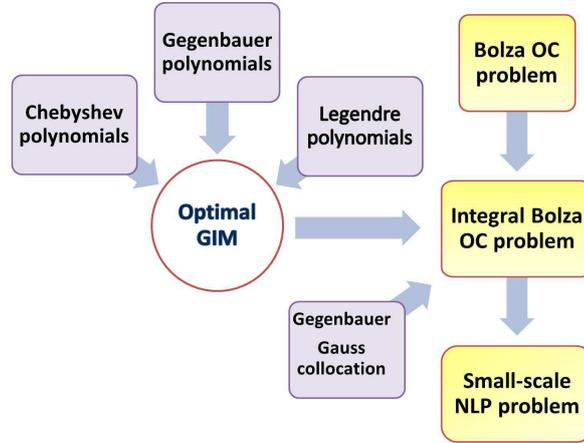


Figure 1.4: The GIM for a continuous-time Bolza OC problem.

problem with considerably lower-dimension than those obtained by other conventional methods. Indeed, this feature is genuinely established using an optimal Gegenbauer quadrature which combines the strengths of the Chebyshev, Legendre, and Gegenbauer polynomials in one unique numerical quadrature using a unified approach. The precise approximations established by the novel quadrature significantly reduces the number of solution points and the Gegenbauer expansion terms needed to accurately represent the state and the control variables. (ii) The GIM discretizes the dynamical system of differential equations using well-conditioned Gegenbauer integration matrices. In contrast, many DOCMs/PS methods presented in the literature apply SDMs known for their ill-conditioning. (iii) An accurate solution can be found using well-developed NLP solvers with no need for an initial guess on the costate or derivation of the necessary conditions. (iv) The GIM takes advantage of the fast exponential convergence typical of spectral methods. (v) The GIM can smoothly represent the state and control dependent constraints. (vi) The GIM offers better convergence behavior than the indirect methods. (vii) The GIM can be quickly used to solve a number of practical trajectory optimization problems since it does not require the derivation of the necessary conditions of optimality. (viii) The GIM solves the CTOCP in the spectral space, which means that once the Gegenbauer spectral coefficients are determined through the NLP solver, the approximation of the state and the control variables can be obtained at any point in the interval of interest. On the other hand, other numerical schemes such as the finite-difference scheme require a further step of interpolation. (ix) One notable advantage of applying the GIM over other DOMs is the high degree of precision offered by the GIM. The rapid convergence rate of the GIM is shown empirically on a wide variety of CTOCPs

## Chapter 1

---

presented later in Chapters 6 and 7 including comparisons with many DOMs presented in the literature. (x) The rapid solution of the CTOCPs enables real-time OC of nonlinear dynamical systems. (xi) Many useful benefits of the GTM are inherited from the application of a DOM endowed with a spectral collocation integration method.

### 1.5 Thesis Overview

This dissertation is divided into eight chapters with five chapters devoted for published articles produced during the Ph.D. candidature. Following this introductory chapter, this dissertation proceeds in the following order: The next chapter presents some mathematical background and basic concepts related to OC theory. In particular, Section 2.1 introduces the CV as the primal basis of OC theory, highlighting the necessary and sufficient conditions required for extremizing functionals subject to boundary conditions. In Section 2.2, we give a fundamental introduction on OC theory. Section 2.2.1 presents some of the popular formulations of CTOCPs, and the necessary conditions of optimality required in these cases. The interesting result of the MP occurs for CTOCPs with input constraints in Section 2.2.1.4. In Section 2.2.2 we review the sufficient conditions of optimality, accentuating on the strengths and weaknesses of DP. Section 2.2.3 introduces the class of DOMs, with a special attention devoted for direct collocation methods (DCMs) in Section 2.2.3.1. In Section 2.3 we present a background on the Gegenbauer polynomials, highlighting some of their useful relations and properties, and introducing some exact formulas for the evaluation of the successive integrations of the Gegenbauer polynomials in terms of themselves. The principles of functions approximations by the Gegenbauer polynomials are briefly highlighted in Section 2.4. The convergence rate of the Gegenbauer collocation methods is discussed in Section 2.5. The concept of the operational matrix of integration is conveniently introduced in Section 2.6. Finally, the idea of solving various mathematical problems such as dynamical systems and OC problems through the optimization of the GIM is briefly presented in Section 2.7.

Chapter 3 presents a published article in *Journal of Computational and Applied Mathematics* titled “Optimal Gegenbauer quadrature over arbitrary integration nodes.” In this chapter, the definite integrations are treated numerically using Gegenbauer quadratures. The presented novel numerical scheme introduces the idea of combining the strengths of the versatile Chebyshev, Legendre, and Gegenbauer polynomials through a unified approach, and using a unique numerical quadrature. In particular, our new vision for constructing the Gegenbauer quadrature efficiently rests upon two fundamental elements: (i) The Gegenbauer polynomial expansions can produce faster convergence rates than both

## Chapter 1

---

the Chebyshev and Legendre polynomials expansions for small/medium range of the spectral expansion terms; (ii) the elegant Chebyshev and Legendre polynomials expansions are optimal in the  $L^\infty$ -norm and  $L^2$ -norm approximations of the smooth functions, respectively, for large-scale number of expansion terms. Therefore, our adopted numerical scheme employs the Gegenbauer polynomials to achieve rapid rates of convergence of the quadrature for the small range of the spectral expansion terms. Moreover, for a large-scale number of the expansion terms, the numerical quadrature possesses the luxury of converging to the optimal Chebyshev and Legendre quadratures in the  $L^\infty$ -norm and  $L^2$ -norm, respectively. The key idea to establish these useful features for a numerical quadrature is to construct the Gegenbauer quadrature through discretizations at some optimal sets of points of the GG type in a certain optimality sense. We show that the Gegenbauer polynomial expansions can produce higher-order approximations to the definite integrals  $\int_{-1}^{x_i} f(x)dx$  of a smooth function  $f(x) \in C^\infty[-1, 1]$  for the small range by minimizing the quadrature error at each integration point  $x_i$  through a pointwise approach. The developed Gegenbauer quadrature allows for approximating integrals for any arbitrary sets of integration nodes. Moreover, exact integrations are obtained for polynomials of any arbitrary degree  $n$  if the number of columns in the developed GIM is greater than or equal to  $n$ . We provide an error formula for the Gegenbauer quadrature, and we address the error bounds and the convergence rates associated with it. Our study manifests that the optimal Gegenbauer quadrature exhibits very rapid convergence rates faster than any finite power of the number of Gegenbauer expansion terms. Our study also demonstrates that there are certain important elements which must be addressed carefully during the construction of the numerical quadrature to establish an efficient and robust numerical scheme. Therefore, we furnished two efficient computational algorithms for optimally constructing a stable and well-behaved Gegenbauer quadrature. We illustrate the high-order approximations of the optimal Gegenbauer quadrature through extensive numerical experiments including comparisons with conventional Chebyshev, Legendre, and Gegenbauer polynomial expansion methods. The theoretical arguments and the experimental results presented in this chapter illustrate the broad applicability of the Gegenbauer quadrature scheme, and its strong addition to the arsenal of numerical quadrature methods.

Chapter 4 presents a published article in *Advances in Computational Mathematics* titled “On the optimization of Gegenbauer operational matrix of integration.” In this chapter we discuss the idea of solving various dynamical systems and OC problems using the GIM by tuning the Gegenbauer parameter  $\alpha$  to achieve better solution approximations. The chapter highlights those methods presented in the literature which apply the Gegenbauer operational matrix of integration for approximating the integral operations, and then recasting the mathematical

## Chapter 1

---

problems into unconstrained/constrained optimization problems. The Gegenbauer parameter  $\alpha$  associated with the Gegenbauer polynomials is then added as an extra unknown variable to be optimized in the resulting optimization problem as an attempt to optimize its value rather than choosing a random value. The chapter focuses exactly on this important topic, and provides a theoretical proof that this operation is indisputably invalid. In particular, we provide a solid mathematical proof demonstrating that optimizing the Gegenbauer operational matrix of integration for the solution of mathematical problems by recasting them into equivalent optimization problems with  $\alpha$  added as an extra optimization variable “violates the discrete Gegenbauer orthonormality relation,” and may in turn produce false solution approximations.

Since an essential stage in the solution of the CTOCPs lies in the efficient discretization of the dynamical system and successfully imposing the boundary conditions, our research work in Chapter 5 is devoted for a comprehensive study on the Gegenbauer spectral solution of general dynamical systems given in many forms, such as differential equations, integral equations, and integro-differential equations. In particular, Chapter 5 presents a published article in *Journal of Computational and Applied Mathematics* titled “Solving boundary value problems, integral, and integro-differential equations using Gegenbauer integration matrices.” In this chapter we introduce a hybrid Gegenbauer integration method for solving BVPs, integral and integro-differential equations. The proposed approach recasts the original problems into their integral formulations, which are then discretized into linear systems of algebraic equations using a hybridization of the GIMs presented in Chapter 3. The resulting linear systems are well-conditioned and can be easily solved using standard linear system solvers. We presented a study on the error bounds of the hybrid technique, and proved the spectral convergence for TPBVPs. We carried out many comparisons with other competitive methods in the recent literature. The hybrid Gegenbauer integration method results in an efficient algorithm, and the spectral accuracy is verified using eight test examples addressing the aforementioned classes of problems. The developed numerical scheme provides a viable alternative to other solution methods when high-order approximations are required using only a relatively small number of solution nodes.

Chapter 6 presents a published article in *The Proceedings of 2012 Australian Control Conference, AUCC 2012*, titled “Solving optimal control problems using a Gegenbauer transcription method.” In this chapter we describe a novel DOCM using GG collocation for solving CTOCPs with nonlinear dynamics, state and control constraints, where the admissible controls are continuous functions. The time domain is mapped onto the interval  $[0, 1]$ , and the dynamical system formulated as a system of ODEs is transformed into its integral formulation through direct integration. The state and the control variables are fully parameterized

## Chapter 1

---

using Gegenbauer expansion series with some unknown Gegenbauer spectral coefficients. The proposed GTM then recasts the performance index, the reduced dynamical system, and the constraints into systems of algebraic equations using the optimal Gegenbauer quadrature developed in Chapter 3. Finally, the GTM transcribes the infinite-dimensional OC problem into a parameter NLP problem which can be solved in the spectral space; thus approximating the state and the control variables along the entire time horizon. In this manner, the GTM retains the structure of the original CTOCP, and solves the problem directly for the states and the controls variables. The high precision and the spectral convergence of the discrete solutions are verified through two OC test problems with nonlinear dynamics and some inequality constraints. In particular, the numerical test problems address the problem of finding the best path for an unmanned aerial vehicle mobilizing in a stationary risk environment. We compared the developed method with the variational technique of Miller et al. (2011)<sup>1</sup>, and we found that the proposed method outperforms the classical variational methods in many aspects such as robustness, simplicity, and accuracy. The results show that the present GTM offers many useful properties and a viable alternative over the available DOMs.

Chapter 7 presents a published article in Journal of Computational and Applied Mathematics titled “Fast, accurate, and small-scale direct trajectory optimization using a Gegenbauer transcription method.” This chapter extends the method introduced in Chapter 6 to deal further with problems including different orders time derivatives of the states by solving the CTOCP directly for the control  $u(t)$  and the highest-order time derivative  $x^{(N)}(t)$ ,  $N \in \mathbb{Z}^+$ . The state vector and its derivatives up to the  $(N - 1)$ th-order derivative can then be stably recovered by successive integration. Moreover, we present our solution method for solving linear-quadratic regulator (LQR) problems as we aim to cover a wider collection of CTOCPs with the concrete aim of comparing the efficiency of the current work with its rivals in the class of direct orthogonal collocation/PS methods. The chapter shows the degree of robustness, simplicity, accuracy, economy in calculations, and speed compared to other conventional methods in the area of computational OC theory. Moreover, the chapter signifies the very important advantage of producing very small-scale dimensional NLP problems, which signals the great gap between the present method and other traditional methods. The advantages of the proposed direct GTM over other traditional discretization methods are shown through four well-studied OC test examples. The present work is a major breakthrough in the area of computational OC theory as it delivers significantly accurate solutions using considerably small numbers of collocation

---

<sup>1</sup>The conference article (Miller et al., 2011) was later updated and published in Journal of Computer and Systems Sciences International as (Andreev et al., 2012).

## Chapter 1

---

points, states and controls expansion terms.

Finally, Chapter 8 presents some concluding remarks on the works achieved in this dissertation including a discussion of promising future research directions.

*Chapter 2*

*Preliminary Mathematical  
Background*

## Chapter 2

# Preliminary Mathematical Background

The foundations of OC theory are grounded on many mathematical subjects such as optimization, standard and variational calculus, linear and nonlinear algebra, approximation theory, numerical analysis, functional analysis, and so on. Clearly, one can discuss only a small fraction of the theory within the scope of a Ph.D. dissertation. However, in this chapter, we shall try to grasp the basic principles and concepts which are useful to our research work.

### 2.1 CV

In this section we briefly discuss the CV as the original foundation of OC theory, and highlight its fundamental theorems. CV is a classical branch of mathematics which dates back to the ancient Greeks, and originated with the works of the great mathematicians of the 17th and 18th centuries. The birth of CV can be directly linked to the Brachistochrone problem (the curve of the shortest descent problem) posed by Galileo Galilei in 1638, and solved later, anonymously, by Johann Bernoulli, Leibniz, Newton, Jacob Bernoulli, Tschirnhaus and L'Hopital (Pytlak, 1999). Although the main concern of the subject is to deal with the optimization of integral functionals, the rigorous developments in this area preceded the development of NLP by many years. CV generalizes ordinary calculus, since its principle objective is to find curves, possibly multidimensional, which optimize certain functionals.

Let  $\mathbf{x}$  be a vector function whose components are continuously differentiable on  $[t_0, t_f]$ , for some fixed real numbers  $t_0, t_f$  with  $t_0 < t_f$ . Also let  $\mathcal{L}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)$  be some functional which is twice differentiable with respect to all of its arguments. One of the simplest variational problems in CV is to find the vector function  $\mathbf{x}^*$

## Chapter 2

---

for which the functional

$$J(\mathbf{x}) = \int_{t_0}^{t_f} \mathcal{L}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t) dt, \quad (2.1)$$

$$\text{subject to } \mathbf{x}(t_0) = \mathbf{x}_0 \text{ and } \mathbf{x}(t_f) = \mathbf{x}_f, \quad (2.2)$$

has a relative extremum. The vector function  $\mathbf{x}^*$  which optimizes the functional  $J(\mathbf{x})$  is said to be an extremal. The fundamental theorem of CV states that if  $\mathbf{x}^*$  is an extremal, then the variation of  $J$  must vanish on  $\mathbf{x}^*$ ; i.e.

$$\delta J(\mathbf{x}^*, \delta \mathbf{x}) = 0 \quad \text{for all admissible } \delta \mathbf{x}. \quad (2.3)$$

The Fundamental theorem plays an essential role in CV, and its application leads to the necessary condition for  $\mathbf{x}^*$  to be an extremal given by

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) - \frac{d}{dt} \left[ \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{x}}}(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) \right] = \mathbf{0} \quad \forall t \in [t_0, t_f], \quad (2.4)$$

together with the boundary conditions  $\mathbf{x}(t_0) = \mathbf{x}_0; \mathbf{x}(t_f) = \mathbf{x}_f$ . The necessary optimality condition (2.4) is known as the “Euler-Lagrange” equation, and it provides a way to solve for functions which extremize a given cost functional. Therefore, it can be considered as the generalization of the condition  $f'(x) = 0$ , for a local extremum of a real variable function  $f(x)$  in standard calculus to the problems of functional analysis, where  $f'(x)$  is the derivative function of  $f(x)$ . Euler-Lagrange equation is generally a nonlinear TPBVP, which gives the condition for a stationarity of a given cost functional. However, this TPBVP usually presents a formidable challenge to be solved both analytically and numerically. It is interesting to know that Legendre (1752–1833) found the additional necessary condition for a minimum by looking at the second variation of  $J$ . His sufficient condition for a minimum asserts that the Hessian matrix must be nonnegative definite, i.e.

$$\nabla_{\dot{\mathbf{x}}, \dot{\mathbf{x}}}^2 \mathcal{L}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t) \geq \mathbf{0}, \quad (2.5)$$

where

$$\nabla_{\dot{\mathbf{x}}, \dot{\mathbf{x}}}^2 \mathcal{L}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t) = \left( \frac{\partial^2 \mathcal{L}}{\partial \dot{x}_i \partial \dot{x}_j}(\mathbf{x}(t), \dot{\mathbf{x}}(t), t) \right)_{1 \leq i, j \leq n}. \quad (2.6)$$

If the Hessian matrix is strictly positive definite, then the condition is said to be a strong Legendre condition.

## 2.2 OC Theory

The distinctive feature of CV is that the optimization of the integral functionals takes place in the space of “all curves.” In contrast, OC problems involve the

## Chapter 2

---

optimization of functionals over a set  $\mathbb{S}$  of curves characterized by some dynamical constraints. Therefore OC theory is considered the natural extension of CV to the problems governed by dynamical system constraints. Historically, this natural extension of OC theory to the CV for systems characterized by ODEs emerged from the aspiration to take various constraints into account (Butkovsky et al., 1968).

OC theory provides the mathematical tool for the construction of efficient control strategies for real life systems. In this framework, one considers a dynamical model evolving over time from one initial state into another, a description of the control mechanism, and a performance criterion defining the objective and the cost of the control action. An OC problem is then formulated as the optimization of the objective function subject to the constraints of the modeling equations (Borzì and Von Winckel, 2009). Hence in its basic form, an OC problem is a set of differential equations describing the paths of the control variables which optimize the cost functional. The main analytical methods for solving OC problems are based upon the MP, and upon the principle of optimality due to Bellman. The computational methods for solving this class of problems have gone through a revolution in the last 25 years, since the introduction of DOCMs in 1988 by Vlassenbroeck and Dooren, and the direct PS methods in 1995 by Elnagar et al., respectively.

According to the type of the system states and controls, the dynamical systems can be conveniently divided into two categories: (i) Continuous dynamical systems, where the control and state variables are functions of a continuous independent variable, usually time or distance; (ii) discrete dynamical systems, where the independent variable changes in discrete increments. An OC problem characterized by continuous components in time such as continuous dynamical system, cost functional, and constraints is called a CTOCP. In the following, we shall describe the mathematical formulation of CTOCPs.

### 2.2.1 Formulation of CTOCPs

There are various formulations for CTOCPs, but generally, a CTOCP requires five essential elements to be well-formulated. These ingredients are: (i) A mathematical model of the system to be controlled, (ii) a specification of the cost function, (iii) a specification of all boundary conditions on the states, (iv) a specification of all the constraints to be satisfied by the states and the controls; (v) a statement of what variables are free; cf. Figure 2.1. In the following, we shall describe some important cases for the mathematical formulations of CTOCPs:

## Chapter 2

---

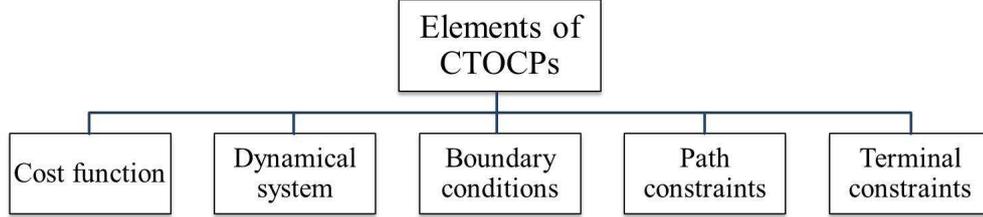


Figure 2.1: The elements of CTOCPs

### 2.2.1.1 Case with Fixed Final Time and No Terminal or Path Constraints

A simple CTOCP is a one with fixed final time and no path or terminal constraints on the states or the control variables. Mathematically this problem can be stated as follows: Find the control vector trajectory  $\mathbf{u} : [t_0, t_f] \mapsto \mathbb{R}^m$  which minimizes the performance index

$$J(\mathbf{u}(t)) = \Phi(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} \mathcal{L}(\mathbf{x}(t), \mathbf{u}(t), t) dt, \quad (2.7a)$$

$$\text{subject to } \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t), \quad (2.7b)$$

$$\mathbf{x}(t_0) = \mathbf{x}_0, \quad (2.7c)$$

where  $[t_0, t_f]$  is the time interval of interest,  $t_f$  is the final/terminal time,  $\mathbf{x} : [t_0, t_f] \mapsto \mathbb{R}^n$  is the state vector,  $\dot{\mathbf{x}} : [t_0, t_f] \mapsto \mathbb{R}^n$  is the vector of first order time derivatives of the states,  $\Phi : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}$  is the terminal cost function,  $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \mapsto \mathbb{R}$  is the Lagrangian function,  $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \mapsto \mathbb{R}^n$  is a vector field where each component  $f_i$  is continuously differentiable with respect to  $\mathbf{x}$  and is continuous with respect to  $\mathbf{u}$  (Bertsekas, 2005). The functions  $\Phi$  and  $\mathcal{L}$  are also continuously differentiable with respect to  $\mathbf{x}$  and  $\mathcal{L}$  is continuous with respect to  $\mathbf{u}$ . Equations (2.7b) and (2.7c) represent the dynamics of the system and its initial state condition. Here we assume that for any admissible control trajectory  $\mathbf{u}(t)$ , the system of differential equations (2.7b) provided with the initial condition (2.7c) has a unique solution denoted by  $\mathbf{x}(t)$ , and is called the corresponding state trajectory. A solution  $\mathbf{u}^*(t)$  of this problem is called an OC. The corresponding curve  $\mathbf{x}^*(t)$  is called the optimal state trajectory; the pair  $(\mathbf{x}^*(t), \mathbf{u}^*(t))$  is usually referred to as the optimal pair. With this notation, the CTOCP (2.7) can be defined as the problem of determining the input  $\mathbf{u}^*(t)$  on the time interval  $[t_0, t_f]$  which drives the plant (2.7b) provided with the initial condition (2.7c) along a trajectory  $\mathbf{x}^*(t)$  such that the cost function (2.7a) is minimized (Lewis and Syrmos, 1995). The problem as defined above is known as the Bolza problem if both  $\Phi$  and  $\mathcal{L}$  are non-zeros. The problem is called the Mayer problem if  $\mathcal{L} = 0$ , and it is known as the Lagrange problem if  $\Phi = 0$ .

## Chapter 2

---

### 2.2.1.2 The Necessary Optimality Conditions Using the Variational Approach

CV is a fundamental technique for defining the extremal solutions to OC problems, and provides the first-order necessary conditions for the optimal solution of the CTOCP (2.7). The idea here is to adjoin the constraints to the performance index using a time-varying Lagrange multiplier vector function  $\boldsymbol{\lambda} : [t_0, t_f] \mapsto \mathbb{R}^n$  to construct the augmented performance index  $J_a$ :

$$J_a = \Phi(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} (\mathcal{L}(\mathbf{x}(t), \mathbf{u}(t), t) + \boldsymbol{\lambda}^T(t)(\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) - \dot{\mathbf{x}})) dt. \quad (2.8)$$

The elements of the vector function  $\boldsymbol{\lambda}$  are also known as the costate/adjoint variables. These variables can be interpreted as the Lagrange multipliers associated with the state equations. To simplify the notation, let us introduce the Hamiltonian function  $H$  as follows:

$$H(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\lambda}(t), t) = \mathcal{L}(\mathbf{x}(t), \mathbf{u}(t), t) + \boldsymbol{\lambda}^T(t)\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t), \quad (2.9)$$

such that  $J_a$  can be rewritten as:

$$J_a = \Phi(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} (H(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\lambda}(t), t) - \boldsymbol{\lambda}^T(t)\dot{\mathbf{x}}) dt. \quad (2.10)$$

An infinitesimal variation  $\delta\mathbf{u}(t)$  in the control history produces variations in the state history denoted by  $\delta\mathbf{x}(t)$ , and a variation in the performance index denoted by  $\delta J_a$  such that

$$\delta J_a = \left[ \left( \frac{\partial \Phi}{\partial \mathbf{x}} - \boldsymbol{\lambda}^T \right) \delta \mathbf{x} \right]_{t=t_f} + [\boldsymbol{\lambda}^T \delta \mathbf{x}]_{t=t_0} + \int_{t_0}^{t_f} \left( \left( \frac{\partial H}{\partial \mathbf{x}} + \dot{\boldsymbol{\lambda}}^T \right) \delta \mathbf{x} + \left( \frac{\partial H}{\partial \mathbf{u}} \right) \delta \mathbf{u} \right) dt. \quad (2.11)$$

Since the Lagrange multipliers are arbitrary, their values can be chosen so that the coefficients of  $\delta\mathbf{x}(t)$  and  $\delta\mathbf{x}(t_f)$  are equal to zero. That is, we can set

$$\dot{\boldsymbol{\lambda}}^T(t) = -\frac{\partial H}{\partial \mathbf{x}}; \quad (2.12)$$

$$\boldsymbol{\lambda}^T(t_f) = \frac{\partial \Phi}{\partial \mathbf{x}} \Big|_{t=t_f}. \quad (2.13)$$

Hence the variation in the augmented performance index is given by:

$$\delta J_a = \int_{t_0}^{t_f} \left( \frac{\partial H}{\partial \mathbf{u}} \right) \delta \mathbf{u} dt, \quad (2.14)$$

## Chapter 2

---

assuming that the initial state is fixed. For  $\mathbf{u}(t)$  to be an extremal, it is necessary that  $\delta J_a = 0$ . This gives the stationarity condition

$$\frac{\partial H}{\partial \mathbf{u}} = 0. \quad (2.15)$$

Equations (2.7b), (2.7c), (2.12), (2.13), and (2.15) are the first-order necessary conditions for a minimum of  $J$ . Equation (2.12) is known as the costate differential equation. Equations (2.7c) and (2.13) represent the boundary/transversality conditions. These necessary optimality conditions define a TPBVP, which may be solved for the analytical solutions of special types of OC problems. However, obtaining closed form solutions is generally out of reach; therefore, these equations are frequently used in defining numerical algorithms and methods known as the IOMs to search for the sought solutions. In these methods, the state equations are solved forwards in time, while the costate equations are solved backwards in time, since the boundary conditions are split.

### 2.2.1.3 Case with Terminal Constraints

Consider the CTOCP (2.7) subject to an additional set of terminal constraints of the form:

$$\boldsymbol{\psi}(\mathbf{x}(t_f), t_f) = \mathbf{0}, \quad (2.16)$$

where  $\boldsymbol{\psi} : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^{n_\psi}$  is a vector function. It can be shown through variational analysis (Lewis and Syrmos, 1995) that the necessary conditions for a minimum of  $J$  are (2.7b), (2.7c), (2.12), (2.15) and the following terminal condition:

$$\left( \frac{\partial \Phi^T}{\partial \mathbf{x}} + \frac{\partial \boldsymbol{\psi}^T}{\partial \mathbf{x}} \boldsymbol{\nu} - \boldsymbol{\lambda} \right)^T \bigg|_{t_f} \delta \mathbf{x}(t_f) + \left( \frac{\partial \Phi}{\partial t} + \frac{\partial \boldsymbol{\psi}^T}{\partial t} \boldsymbol{\nu} + H \right) \bigg|_{t_f} \delta t_f = 0, \quad (2.17)$$

where  $\boldsymbol{\nu} \in \mathbb{R}^{n_\psi}$  is the Lagrange multiplier associated with the terminal constraint,  $\delta t_f$  is the variation of the final time;  $\delta \mathbf{x}(t_f)$  is the variation of the final state. Note here that if the final time is fixed, then  $\delta t_f = 0$  and the second term vanishes. Also, if the terminal constraint is such that the  $j$ th element of  $\mathbf{x}$  is fixed at the final time, then the  $j$ th element of  $\delta \mathbf{x}(t_f)$  vanishes.

### 2.2.1.4 Case with Input Constraints– The MP

Physically realizable controls generally have magnitude limitations. Moreover, admissible states are constrained by certain boundaries due to certain measures such as safety, structural restrictions, etc. Therefore, state and control constraints commonly occur in realistic dynamical systems. Pontryagin and coworkers (Boltyanskii et al., 1956; Pontryagin et al., 1962) established the MP, which

## Chapter 2

---

provides the necessary conditions of optimality in the presence of constraints on the states or input controls. The MP is one of the most important results in OC theory. Its main idea is to transfer the problem of finding the input  $\mathbf{u}(t)$  which minimizes the cost function subject to the given constraints, to the problem of minimizing the Hamiltonian function with respect to it (Grigorenko, 2006). Historically, the principle was first applied to the minimum-time problems ( $\Phi = 0; \mathcal{L} = 1$ ) where the input control is constrained. Moreover, it can be considered an extension of Weierstrass necessary condition to cases where the control functions are bounded.

Let the input vector  $\mathbf{u} \in \mathbb{U}$ , where  $\mathbb{U}$  is the set of all permissible controls. It was shown by Pontryagin and co-workers (Boltyanskii et al., 1956; Pontryagin et al., 1962) that in this case, the necessary conditions (2.7b), (2.7c), (2.12), and (2.13) still hold, but the stationarity condition (2.15) has to be replaced by:

$$H(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t), t) \leq H(\mathbf{x}^*(t), \mathbf{u}(t), \boldsymbol{\lambda}^*(t), t) \quad \forall \mathbf{u} \in \mathbb{U}, t \in [t_0, t_f], \quad (2.18)$$

where  $\boldsymbol{\lambda}^*$  is the optimal costate trajectory. Hence for a minimum of the cost function in the case of input constraints, the Hamiltonian must be minimized over all admissible  $\mathbf{u}$  for optimal values of the state and costate variables. The OC obtained through the MP is called an “open-loop control,” since it is a function of the time  $t$  only, and one applies this control function thereafter with no further observation of the state of the system. It is important to note that the MP provides the necessary conditions of optimality, but it is generally not sufficient for any control trajectory satisfying these conditions to be truly optimal. That is, using the MP alone, one is often not able to conclude that a trajectory is optimal. The MP is deemed sufficient in the trivial cases when there exists only one control trajectory satisfying the MP conditions, or when all control trajectories satisfying these conditions have equal cost.

The formidable necessary conditions of optimality (2.7b), (2.7c), (2.12), (2.13), and (2.18) lead to a generally nonlinear TPBVP with a mixed initial/terminal boundary conditions for the system state and its conjugate costate. This reduced TPBVP must be solved to obtain the OC law  $\mathbf{u}$ . Generally, this is a very difficult task both analytically and computationally, since the TPBVP is known to be very unstable, and the MP does not give any information on the initial values of the costates (Liu, 2011). Due to the instability of the TPBVP, determining an OC law is possible only for systems with a “perfect” model, and at the cost of losing beneficial properties such as robustness with respect to disturbances and modeling uncertainties (Lin, 2011).

A special case where the solution can be obtained in closed loop form is the LQR, where the plant is linear and the performance index is a quadratic form. However, in general, the necessary conditions of optimality provided by the MP are intractable for analytical or closed form expressions of the control law. Even

## Chapter 2

---

in the mild case of a LQR, the OC law is usually determined by solving a matrix differential equation of the fierce Riccati type. In fact, except for very special cases, it is well-known that obtaining analytical solutions of Riccati differential equations is usually out of reach, and their numerical solution is still undoubtedly a “daunting” task (Anderson and Moore, 2007; Deshpande and Agashe, 2011; Reid, 1972).

### 2.2.1.5 Special Cases— Some Results Due to the MP

In this section, we briefly state, without proof, some results due to the MP for various OC problems:

- The MP can still be applied if the control  $\mathbf{u}$  is not bounded within a certain constraint region. In this case, for  $\mathbf{u}^*(t)$  to minimize the Hamiltonian, Equation (2.15) must hold and the matrix

$$\frac{\partial^2 H}{\partial \mathbf{u}^2}(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t), t), \quad (2.19)$$

must be positive definite.

- If  $t_f$  is fixed, and the Hamiltonian does not depend explicitly on time, then the Hamiltonian must be constant when evaluated on an extremal trajectory, i.e.  $H(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) = \text{constant} \forall t \in [t_0, t_f]$ .
- If  $t_f$  is free, and the Hamiltonian does not explicitly depend on time, then the Hamiltonian must be identically zero when evaluated on an extremal trajectory, i.e.  $H(\mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}^*(t)) = 0 \forall t \in [t_0, t_f]$ .
- If  $\mathbf{x}(t_0)$  is free, then  $\boldsymbol{\lambda}(t_0) = 0$  holds and represents an extra boundary condition for the adjoint equation (Bertsekas, 2005).
- If  $\mathcal{L}(\mathbf{x}(t_0))$  represents a cost on  $\mathbf{x}(t_0)$ , then the boundary condition becomes  $\boldsymbol{\lambda}(t_0) = -\nabla \mathcal{L}(\mathbf{x}^*(t_0))$ .

Other cases may occur due to the existence of some path constraints of the form  $\mathbf{c}(\mathbf{x}(t), \mathbf{u}(t), t) \leq 0$ , such that  $\mathbf{c} : \mathbb{R}^n \times \mathbb{R}^m \times [t_0, t_f] \mapsto \mathbb{R}^{n_c}$ , or equality constraints at some intermediate points in time, or some singular arcs in the solutions of the OC problems, etc.; cf. (Bertsekas, 2005; Betts, 2009; Bryson and Ho, 1975; Subchan and Zbikowski, 2009).

## Chapter 2

---

### 2.2.1.6 General Mathematical Formulation of CTOCPs

One of the most general mathematical formulations of CTOCPs is the one which includes the five essential elements described in Figure 2.1. In particular, a common general CTOCP formulation has free final time  $t_f$ , nonlinear dynamics with mixed boundary conditions, mixed path and terminal constraints, and it can be described as follows: Find the control vector trajectory  $\mathbf{u} : [t_0, t_f] \subset \mathbb{R} \rightarrow \mathbb{R}^m$  which minimizes the general cost function:

$$J(\mathbf{u}(t)) = \Phi(\mathbf{x}(t_0), t_0, \mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} \mathcal{L}(\mathbf{x}(t), \mathbf{u}(t), t) dt, \quad (2.20a)$$

subject to the nonlinear plant (2.7b), and the following constraints:

$$\phi(\mathbf{x}(t_0), t_0, \mathbf{x}(t_f), t_f) = \mathbf{0}, \quad (2.20b)$$

$$\boldsymbol{\psi}(\mathbf{x}(t), \mathbf{u}(t), t) \leq \mathbf{0}, \quad t \in [t_0, t_f], \quad (2.20c)$$

$$\mathbf{h}(\mathbf{x}(t_f), \mathbf{u}(t_f), t_f) \leq \mathbf{0}, \quad (2.20d)$$

where  $\phi : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^{n_\phi}$ , for some  $n_\phi \in \mathbb{Z}^+$ , represents the state boundary conditions in their most general form,  $\boldsymbol{\psi} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^{n_\psi}$  is a vector field, where each component  $\psi_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}$  represents a state and control inequality constraint for each  $i = 1, \dots, n_\psi$ ;  $n_\psi \in \mathbb{Z}^+$ ;  $\mathbf{h} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}$  is a vector field representing a state and control terminal inequality constraint. The first-order necessary conditions for the optimal solution of similar CTOCPs derived through the variational techniques can be found in (Benson, 2004; Darby, 2011; Garg, 2011).

### 2.2.2 DP– Sufficient Conditions of Optimality

DP constitutes the first extension of the CV to problems with inputs (Polak, 1973). The subject is considered an alternative to the variational approach in OC theory, and is a means by which candidate OCs can be verified optimal. DP was proposed by Bellman in 1953, and represents an extension to the Hamilton-Jacobi theory. It is concerned mainly with the families of extremal paths which meet specified terminal conditions (Bryson, 1996). Bellman realized that ‘*an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision*’ (Kirk, 2004). In this manner, DP makes the direct search feasible by considering only the controls which satisfy the principle of optimality rather than searching among the set of all admissible controls which yield admissible trajectories. Hence the main idea of the optimality principle is to determine the OC by limiting the number of potential OC strategies which must

## Chapter 2

---

be investigated. Moreover, the optimal strategy must be developed backward in time, i.e. determined by working backward from the final time.

Bellman translated the principle of optimality into a conceptual method for solving dynamic optimization problems. The framework of the method can be outlined by introducing the optimal cost-to-go function (value function)  $J^*(\mathbf{x}(t), t)$ , as follows:

$$J^*(\mathbf{x}(t), t) = \min_{\mathbf{u}(\cdot)} \left\{ \Phi(\mathbf{x}(t_f), t_f) + \int_t^{t_f} \mathcal{L}(\mathbf{x}(t), \mathbf{u}(t), t) dt \mid \mathbf{x}(t) = \mathbf{x} \right\}. \quad (2.21)$$

Equation (2.21) gives the optimal remaining cost when the dynamical system is in state  $\mathbf{x}$  at time  $t$ . It can be shown that Bellman's principle of optimality applied to the CTOCP (2.7) leads to the HJB equation, which can be written in terms of the optimal cost-to-go function  $J^*(\mathbf{x}(t), t)$  in the following form:

$$-\frac{\partial J^*}{\partial t}(\mathbf{x}(t), t) = \min_{\mathbf{u}(t)} \left( \mathcal{L}(\mathbf{x}(t), \mathbf{u}(t), t) + \frac{\partial J^*}{\partial \mathbf{x}}(\mathbf{x}(t), t) \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) \right) \quad \forall t, \mathbf{x}, \quad (2.22)$$

with the boundary condition  $J^*(\mathbf{x}(t_f), t_f) = \Phi(\mathbf{x}(t_f), t_f)$ , assuming that  $J^*$  is continuously differentiable in its arguments. Equation (2.22) is a PDE which must be satisfied for all time-state pairs  $(t, \mathbf{x})$  by the optimal cost-to-go function  $J^*(\mathbf{x}(t), t)$ . If we can solve for  $J^*$ , then the OC  $\mathbf{u}^*$  which achieves the minimum cost can be found from it.

It is noteworthy to mention that the HJB Equation (2.22) is a necessary and sufficient condition for an optimum if it is solved over the whole state space. Moreover, the OC  $\mathbf{u}^*(\mathbf{x}, t)$  associated with the optimal cost-to-go function  $J^*(\mathbf{x}(t), t)$  is known as the ‘‘closed-loop control,’’ since it is a feedback on the current state  $\mathbf{x}$  and the time  $t$ . This is why DP may be called ‘‘nonlinear optimal feedback control’’ (Bryson, 1996). Hence a distinctive feature between the MP and DP is that the former produces an open-loop control, while the latter yields a closed-loop control. Moreover, the co-states of the MP are valid only for a particular initial state, while the value function of DP covers the whole state space (Rungger and Stursberg, 2011). Furthermore, DP provides the sufficient and necessary conditions for optimality while the MP offers only the set of necessary conditions of optimality. In other words, while the OC obtained through CV and the MP is a local OC in general, DP ensures that the OC obtained through the HJB equation is actually the global OC law by directly comparing the performance index values associated with all of the OC law candidates. The solution obtained through DP is more attractive indeed, whereas in real systems, there will usually be random perturbations which are not accounted for in the mathematical model. These random disturbances can be automatically corrected using a feedback law. Besides this useful feature, the existence of the state and control constraints simplifies the

## Chapter 2

---

application of DP, since they reduce the range of the values to be searched and thereby simplify the solution. In contrast, the presence of the state and control constraints generally complicates the application of the variational techniques.

### 2.2.2.1 The Curse of Dimensionality– A Fatal Drawback Associated with the DP Solution Method

It is well-known that the HJB equation is hard to solve, and rarely admits analytical solutions to OC problems, especially for nonlinear systems. In fact, explicit solution exists in cases like the LQR, but in general, numerical methods must be employed. Finding the value function from the HJB PDE is usually carried out backwards in time, starting from  $t = t_f$  and ending at  $t = t_0$ .

The aforementioned features of DP described in Section 2.2.2 seem to suggest that it is the method of choice for solving OC problems. In fact, despite all of the useful advantages of DP, there is a fatal drawback which greatly limits its application, especially for high-dimensional systems: All numerical methods for solving the HJB equation are computationally intense, and subject to the so-called “curse of dimensionality” (Lawton et al., 1999; Murray et al., 2002). In particular, for high-dimensional systems, the solution of the HJB equation demands a huge number of computations and storage space to the extent that the number of high-speed storage locations becomes prohibitive (Kirk, 2004; Lin, 2011). To make matters worse, most of the numerical methods for solving the HJB equation are usually incapable of producing stable and robust OC solutions. These limitations are also true for Bellman’s equation, which is a functional recurrence relation representing the discrete analogy of the HJB equation. The interested reader may consult (Bellman, 2003; Bertsekas, 2007; Kirk, 2004) for further information about DP.

### 2.2.3 DOMs

The topic of DOMs is very large, and various methods have been presented in the literature and have been actively used. DOMs retain the structure of the original infinite-dimensional CTOCP and transcribe it directly into a finite-dimensional parameter NLP problem through the discretization of the original problem in time and performing some parameterization of the control and/or state vectors. There are three main types of DOMs, namely: (i) DOMs based on parameterizing the controls only (partial parameterization), (ii) DOMs based on parameterizing the states only (partial parameterization), and (iii) DOMs based on parameterizing both the states and the controls (full parameterization).

In the first type of DOMs, the parameterization of the control profile simply means specifying the control input function through some parameters which are

## Chapter 2

---

allowed to take on values in a given specified range. The parameterization of the control can take many forms. Generally, one can choose  $\mathbf{u} = \mathbf{u}(\boldsymbol{\alpha}, t)$ , for some parameter vector  $\boldsymbol{\alpha} \in \mathbb{R}^l; l \in \mathbb{Z}^+$ . The most successful parameterization strategies for the control in the literature are based on the trial function expansions, which are widely used in solving OC problems as well as in many other areas. In fact, the parameterization of the control profile

$$u(t) = \sum_{i=0}^N a_i \phi_i(t), \quad (2.23)$$

by expanding in functions of time has been originally introduced by Rosenbrock and Storey (1966), where  $a_i$  are some finitely many unknown coefficients, and  $\phi_i(t)$  are some arbitrary functions, which depend on the underlying problem. In this manner, the OC problem is transformed into a static optimization problem in the coefficients  $a_i$ . The direct single-shooting method is one of the earliest methods which belong to the first type of DOMs. In a typical direct single-shooting method, only the controls are to be parameterized, and the differential constraints must then be integrated over the entire time domain using a numerical integration scheme such as Euler's method, Heun method, or Runge-Kutta method, etc., where the states are obtained recursively; cf. (Betts, 2009). One major drawback of this class of methods is their severe instability in many cases, which limit their applications.

In the second type of DOMs, only the state variables are parameterized. The control vector is then obtained from the system state equations as a function of the approximated state vector. The OC problem in this manner is transformed into a parameter optimization problem; cf. (Nagurka and Wang, 1993; Sirisena and Chou, 1981).

In the third type of DOMs, both the states and the controls are fully parameterized as follows:

$$\mathbf{x} = \mathcal{F}_1(\boldsymbol{\alpha}, t), \quad (2.24a)$$

$$\mathbf{u} = \mathcal{F}_2(\boldsymbol{\beta}, t), \quad (2.24b)$$

respectively, for some prescribed vector functions  $\mathcal{F}_1; \mathcal{F}_2$ , and parameter vectors  $\boldsymbol{\alpha} \in \mathbb{R}^{n_\alpha}; \boldsymbol{\beta} \in \mathbb{R}^{n_\beta}$ . In this manner, the OC problem can be easily transformed into a parameter NLP problem with the unknown optimization vectors  $\boldsymbol{\alpha}; \boldsymbol{\beta}$  through the discretization of the cost function, the dynamics, and the constraints. Direct multiple-shooting methods are examples of DOMs of the third type. These methods are typically carried out by dividing the solution domain into several subintervals as an attempt to reduce the instability effect encountered during the implementation of the direct single-shooting method by shooting over shorter

## Chapter 2

---

steps; cf. (Betts, 2009; Bock and Plitt, 1984; Fraser-Andrews, 1999; Schwartz, 1996). Although direct multiple-shooting methods are considered more stable than direct single-shooting methods, the expensive calculations required by the numerical integrator in these methods is a major drawback in their applications. DCMs are other examples of DOMs of the third type. These methods usually present more efficiency and robustness over direct shooting methods. In fact, whereas shooting methods rely on a separate integrator to solve the dynamical system, DCMs include the integration in the resulting optimization problem as constraints. That is, the dynamical system is converted into a system of algebraic equations relating the consecutive states and controls as constraints. Therefore, implicit integration is efficiently carried out as the set of state equations is solved as a part of the optimization problem. It has been proven that the use of implicit integration combined with modern NLP packages is an effective technique for trajectory optimization (Paris et al., 2006).

In DCMs, where implicit integration is efficiently applied, a general OC problem is discretized at a certain points set  $\{t_i\}_{i=0}^N$  and transformed into a constrained NLP problem, which can be stated as follows: Find a decision vector  $\mathbf{y} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^T \in \mathbb{R}^{n_\alpha + n_\beta}$  which minimizes

$$J(\mathbf{y}), \quad (2.25a)$$

subject to

$$\mathbf{h}(\mathbf{y}) = \mathbf{0}, \quad (2.25b)$$

$$\mathbf{g}(\mathbf{y}) \leq \mathbf{0}, \quad (2.25c)$$

where  $J : \mathbb{R}^{n_\alpha + n_\beta} \rightarrow \mathbb{R}$  is a differentiable scalar function,  $\mathbf{h} : \mathbb{R}^{n_\alpha + n_\beta} \rightarrow \mathbb{R}^{n_h}$ ,  $n_h \in \mathbb{Z}^+$ ;  $\mathbf{g} : \mathbb{R}^{n_\alpha + n_\beta} \rightarrow \mathbb{R}^{n_g}$ ,  $n_g \in \mathbb{Z}^+$  are differentiable vector functions. The NLP problem (2.25) can then be solved for  $\mathbf{y}$  using well-developed optimization software. Once the decision vector  $\mathbf{y}$  is found, the states and the controls can be evaluated directly at any time history in the solution domain using Equations (2.24).

Another approach for implementing DCMs is to formulate the NLP in terms of the state vectors  $\mathbf{x}_i = \mathbf{x}(t_i)$  and the control vectors  $\mathbf{u}_i = \mathbf{u}(t_i)$  instead of the parameter vectors  $\boldsymbol{\alpha}; \boldsymbol{\beta}$ . In this manner, the NLP problem can be solved in the physical space instead of the parameter space (the space of the parameter vectors). The sought decision vector in this case is  $\mathbf{y} = (\mathbf{X}, \mathbf{U})^T$ , where

$$\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_N)^T; \quad (2.26)$$

$$\mathbf{U} = (\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_N)^T. \quad (2.27)$$

Theoretically, it can be shown that the above two approaches are equivalent, i.e. one can determine the physical values of the states and the controls if the parameter vectors are known, and vice-versa.

## Chapter 2

---

Many DOMs and IOMs in the literature are based on the control and/or state parameterizations (or their physical discretizations) techniques; cf. (Betts, 2001; Elnagar et al., 1995; Enright and Conway, 1992; Fahroo and Ross, 2001, 2002; Garg et al., 2011b; Goh and Teo, 1988; Hager, 2000; Hargraves and Paris, 1987; Hicks and Ray, 1971; Huntington, 2007; Sirisena, 1973; Sirisena and Tan, 1974; Sirisena and Chou, 1981; Stryk, 1993; Teo and Womersley, 1983; Teo and Wong, 1992; Vlassenbroeck and Dooren, 1988; Williams, 2004; Wong et al., 1985). The size of the resulting NLP problem depends on the parameterization method. Although partial parameterization (parameterizing only the controls or the states) results in a smaller optimization problem than full parameterization (parameterizing both the controls and the states), the latter schemes are generally more stable and their resulting optimization problems are usually well-conditioned.

### 2.2.3.1 DCMs

DCMs are based on the direct application of the standard collocation methods for the solution of OC problems. Collocation methods form one of the three main classes of the popular spectral methods, namely, the Galerkin methods, the Lanczos tau-methods, and the collocation/PS methods. Spectral methods were largely developed in the 1970s for solving PDEs arising in fluid dynamics and meteorology (Canuto et al., 1988), and entered the mainstream of scientific computation in the 1990s (Canuto et al., 2006). These methods gained much attention from the researchers that they became ‘one of the “big three” technologies for the numerical solution of PDEs’ (Trefethen, 2000). Over the last 25 years, spectral methods have emerged as important computational methods for solving complex nonlinear OC problems; cf. (Elnagar et al., 1995; Elnagar and Kazemi, 1998a; Gong et al., 2006a; Jaddu, 2002; Jaddu and Shimemura, 1999; Vlassenbroeck, 1988; Vlassenbroeck and Dooren, 1988; Williams, 2004), and many other articles in the literature of OC theory.

In a typical spectral collocation method, the function  $f(x)$  is approximated by a finite expansion such as

$$f_n(x) = \sum_{k=0}^n c_k \phi_k(x), \quad (2.28)$$

where  $\{\phi_k\}_{k=0}^n$  is a chosen sequence of prescribed globally smooth basis functions. One then proceeds to estimate the coefficients  $\{c_k\}_{k=0}^n$ ; thus approximating  $f(x)$  by a finite sum. When the series (2.28) is substituted into the differential/integral equation

$$L f(x) = g(x), \quad (2.29)$$

## Chapter 2

---

where  $L$  is the operator of the differential or the integral equation, the result is the so-called “residual function” defined by

$$\mathcal{R}(x, c_0, c_1, \dots, c_n) = Lf_n - g. \quad (2.30)$$

The goal is to choose the series coefficients  $\{c_k\}_{k=0}^n$  so that the residual function  $\mathcal{R}$  is minimized. The collocation method demands that the differential/integral equation (2.29) be exactly satisfied at a set of points  $\{x_i\}_{i=0}^N$  within the physical domain known as the “collocation” points. The consequence of this condition is that Equation (2.29) is fulfilled exactly at the collocation points, i.e.  $Lf_n(x_i) = g(x_i), i = 0, \dots, N$ . One can also derive condition (2.30) using the well-known method of weighted residuals, which requires that the residual  $\mathcal{R}$  multiplied with  $(N + 1)$  test functions  $\{\omega_i(x)\}_{i=0}^N$ , and integrated over the solution domain  $\mathcal{D}$  must vanish (Elgindy, 2008), i.e.

$$(\mathcal{R}, \omega_i) = \int_{\mathcal{D}} \omega_i(x) \mathcal{R}(x, c_0, \dots, c_n) dx = 0, \quad i = 0, \dots, N,$$

where “ $(\cdot, \cdot)$ ” is the inner product defined by

$$(f, g) = \int_{\mathcal{D}} f(x)g(x)dx,$$

for any two functions  $f(x); g(x)$ . In collocation methods, the test functions are

$$\omega_i(x) = \delta(x - x_i), \quad i = 0, \dots, N,$$

with  $\delta$  being the Dirac delta function

$$\delta(x) = \begin{cases} 1, & \text{for } x = 0, \\ 0, & \text{otherwise.} \end{cases}$$

The collocation method is known as an orthogonal collocation method if the chosen basis functions are orthogonal. The gained popularity of these methods compared to Galerkin methods and Lanczos tau-methods is largely due to their greater simplicity and computational efficiency. In fact, the collocation methods can solve nonlinear and variable coefficient problems more efficiently than Galerkin, or Lanczos tau approximations (Gottlieb and Orszag, 1977). A thorough discussion on the collocation methods, and spectral methods in general can be found in many useful textbooks and monographs; cf. (Boyd, 2001; Fornberg, 1996; Gottlieb and Orszag, 1977; Trefethen, 2000) for instances, and the references therein.

## Chapter 2

---

### 2.2.3.2 DLCMs

DCMs are conveniently divided into two categories: DLCMs and direct global collocation methods (DGCMs). In the former, either the controls or both the controls and the states are parameterized using piecewise polynomials. Typically fixed low-degree polynomials for the approximation of the state and the control variables are used. The duration time of the optimal trajectory is divided into several subintervals using the collocation points  $\{t_i\}_{i=0}^N$ . The dynamics and the constraints are then imposed at the collocation points in the solution domain (Reeger, 2009). The convergence of the numerical scheme is achieved by increasing the number of segments. To obtain specified solution accuracy, some grid refinement methods increase the number of mesh intervals in regions of the trajectory where the errors are largest (Darby et al., 2011). Examples of DLCMs are the Euler method (Dontchev and Hager, 1997), and the second-order Runge-Kutta method (Dontchev et al., 2000) for the solution of state and control constrained OC problems, respectively. A main drawback in such finite-difference schemes is their algebraic convergence rates in  $N$ , which are typically of  $O(N^{-2})$  or  $O(N^{-4})$  (Weideman and Reddy, 2000).

### 2.2.3.3 DGCMs

In DGCMs, the controls and/or the states are parameterized using global trial functions defined across the entire time interval. The values of the state and control variables are then sought at a certain set of collocation points. The time derivatives of the states in the dynamic equations are approximated by evaluating the derivatives of the global interpolating polynomials. These approximate derivatives are then constrained to be equal to the vector field of the dynamic equations at the set of collocation points. The most successful methods in this area employ global orthogonal polynomials as the trial basis polynomials, in particular, those which belong to the Jacobi family of polynomials; cf. (Benson, 2004; Benson et al., 2006; Darby, 2011; Elnagar et al., 1995; Elnagar, 1997; Fahroo and Ross, 2002, 2008; Garg, 2011; Garg et al., 2011a; Gong et al., 2006a; Williams, 2004). These methods are called DOCMs or direct PS methods, and are considered the biggest technology in the area of OC theory in the last quarter century. Many complex OC problems have been exclusively solved by the direct PS methods using the OTIS FORTRAN software package (Paris and Hargraves, 1996) and the DIDO MATLAB software codes (Ross, 2004). As a result of the considerable success of these methods, NASA applied the Legendre PS method as a problem solving option for their OTIS software package (Gong et al., 2008). Furthermore, the current PSOPT OC C++ software package (Becerra, 2011) uses direct PS methods as a problem solving option, which has been applied recently to help

## Chapter 2

---

design optimal trajectories for the first Brazilian deep space mission to the triple asteroid system 2001 SN263, which is due to be launched in 2016. One substantial advantage of the DOCMs and direct PS methods over other traditional discretization methods is their promise of exponential convergence for smooth problems. This rapid convergence is faster than any polynomial convergence rate (Canuto et al., 1988), exhibiting the so called “spectral accuracy” while providing Eulerian-like simplicity. That is, these methods converge to the solutions faster than  $O(N^{-m})$ , where  $N$  is the number of collocation points, and  $m$  is any finite power value (Rao et al., 2010). Another advantage of DOCMs employing the Jacobi family of polynomials as the basis polynomials appears in the orthogonality properties of these complete and easily evaluated type of polynomials, which provide the fast conversion between the spectral coefficients  $\{a_i\}_{i=0}^N$  and the function values at the set of collocation nodes  $\{t_i\}_{i=0}^N$ . In contrast, unless the states and the controls in (2.24) are well represented in suitable parametric representations, determining the parameter vectors  $\boldsymbol{\alpha}; \boldsymbol{\beta}$  from the physical values of the states and the controls may not be an easy task. In fact, the values of the states and the controls at an intermediate point  $\bar{t} \notin \{t_i\}_{i=0}^N$  in this case may not be readily determined, since obtaining the solution vectors  $\mathbf{x}(\bar{t}); \mathbf{u}(\bar{t})$  may require solving a highly nonlinear algebraic (or transcendental) system of equations, and one may invoke an interpolation method in such cases.

### 2.2.3.4 Choice of the Collocation Points in a DOCM/PS Method

While any set of collocation points can be used in a DOCM or a direct PS method, many theoretical and experimental results found in the literature favor the orthogonal collocation set of points over other choices of collocation points sets such as the equidistant grid. That is, the collocation points are preferably chosen to be the zeros of the orthogonal basis polynomials employed in the series expansion approximations, or the zeros of linear combinations of such polynomials and their derivatives (Huntington, 2007). One major reason for this particular choice of nodal points is to avoid the Runge phenomenon and the divergence of the approximating interpolating polynomial in the interpolation of nonperiodic functions on finite intervals (Fornberg, 1996; Isaacson and Keller, 1994). In fact, it is well-known that one must abandon the equidistant grid, and choose a grid which clusters quadratically as

$$x_j \sim -1 + c(j/N)^2, \quad (2.31)$$

close to the endpoints. In other words, a suitable grid of points must have an asymptotic distribution with a density  $\rho$  such that

$$\rho \sim \frac{N}{\pi\sqrt{1-x^2}}, \quad (2.32)$$

## Chapter 2

---

per unit length, as  $N \rightarrow \infty$  (Trefethen, 2000). In fact, all the zeros of classical orthogonal polynomials and their derivatives satisfy this important condition (Gottlieb and Hesthaven, 2001). Moreover, the best well-known and accurate quadrature rules use these types of collocation points as the quadrature points.

It is noteworthy to mention that depending on the form of the coefficients  $a_i$  in the parameter expansion series (2.23), a DOM using global orthogonal basis/interpolating polynomials is termed a DOCM or a direct PS method. In particular, while the expansion coefficients in a DOCM may assume any value, they are the function values at the collocation points in a direct PS method; cf. (Fahroo and Ross, 2002) for instance. Therefore direct PS methods are considered a subclass of DOCMs. Some other researchers deem the terms “PS” and “orthogonal collocation” identical, and have the same meaning; cf. (Garg et al., 2011b) for instance. In both cases, the expansion series are called the spectral expansion series, and the coefficients are called the spectral coefficients. In this dissertation, we shall follow the former definition.

### 2.2.3.5 The Framework of Solving OC Problems Using DOCMs/PS Methods

In a typical DOCM, the control and/or the state variables are approximated by orthogonal polynomials series expansions over the whole time horizon. The integrals are approximated by quadratures (truncated sums) over a certain set of quadrature points usually of the Gauss type, while the derivatives are approximated by discrete differential operators known as the SDMs. In this manner, the performance index, the system dynamics, the boundary conditions, the state and/or control constraints are all converted into algebraic expressions. Eventually, the original OC problem is discretized and converted into a constrained NLP problem of the form (2.25), where the sought decision vector  $\mathbf{y}$  is a vector of the state and the control variables’ coefficients in the spectral space, or a vector of the state and the control variables’ values at the time collocation nodes in the physical space. The convergence of DOCMs is achieved by increasing the number of collocation points and spectral expansion terms, and the degree of the polynomial approximation (Darby et al., 2011). There are two common traits between the DGCMs and DLCMs in the framework of solving OC problems: (i) The system dynamics and the constraints are enforced/fulfilled pointwise only (locally at the collocation points); (ii) no integration of differential equations is to be carried out, which is very attractive (Pesch, 1994).

## Chapter 2

---

### 2.2.3.6 DGCMs Versus DLCMs

While DLCMs have been applied frequently for solving OC problems, the computational efficiency and the accuracy achieved by these methods are much less than those accomplished by DOCMs for sufficiently differentiable solutions. There are many reasons why DOCMs/PS methods are more competitive than other DLCMs for solving OC problems with smooth solutions. The most two significant features are the simplicity of the discretization procedure and the very precise accuracy (spectral accuracy) of the method. In fact, it has been demonstrated in the literature that the approximated states and controls using DLCMs converge at a much slower rate than the observed convergence rates of global methods. For a desired accuracy, DLCMs require significantly more computational effort as compared to the global methods both in terms of the calculated time and the number of iterations of the NLP solver (Huntington, 2007). Gong et al. (2006a) specifically compared the efficiency of direct global PS methods versus some other DLCMs for solving nonlinear OC problems with smooth solutions. In particular, Table I in (Gong et al., 2006a) shows a comparison between the efficiencies of three discretization methods implementing Euler’s method, the Hermite-Simpson method, and a PS method. The comparison includes the difference between the number of required nodes to achieve threshold accuracy for each method, and the calculation time required by each method to obtain the solutions. The results of the table clearly show the superiority of the PS methods over the DLCMs for solving nonlinear OC problems with smooth solutions. For example, the PS method requires 18 collocation nodes to achieve an error of  $O(10^{-08})$  in 0.326 seconds, while the Hermite-Simpson method requires 70 collocation nodes to obtain an error of  $O(10^{-04})$  in 1.465 seconds. The results further show that Euler’s method requires 500 collocation points to produce an error of  $O(10^{-03})$  in 37.451 seconds! Indeed, DOCMs and direct PS methods do not require extremely large number of variables in their approximations as needed in an Eulerian discretization to obtain “comparable” precisions of solutions (Polak, 2011). Moreover, DOCMs/PS methods can produce significantly small-scale NLP problems, which can be solved quickly, and result in very precise approximations. This property is extremely attractive for control applications as ‘it places real-time computation within easy reach of modern computational power’ (Gong et al., 2007). On the other hand, typical DLCMs applying finite-difference schemes such as Euler or Runge-Kutta methods lead to enormous optimization problems in the number of decision variables and constraints to obtain higher-order approximations, which is very expensive and time-consuming. Hence, it can be clearly seen that the contest between DGCMs and DLCMs for OC problems with smooth solutions is not an even battle, but rather a rout. DGCMs win hands down. Here it is noteworthy to mention that some fundamental results on the feasibility, consistency,

## Chapter 2

---

and convergence of direct PS methods for solving OC problems have been shown in a number of articles; cf. (Gong et al., 2006a, 2009, 2008; Kang et al., 2007, 2008; Ruths et al., 2011). The rate of convergence of direct PS methods has been proven recently by Kang (2009).

### 2.3 Gegenbauer Polynomials

The theory of Gegenbauer polynomial approximation has received considerable attention in recent decades (Archibald et al., 2003; Area et al., 2004; Ben-yu, 1998; Doha, 1990, 2002; Doha and Abd-Elhameed, 2009; El-Hawary et al., 2000, 2003; Elbarbary, 2006; Gelb, 2004; Gelb and Jackiewicz, 2005; Gottlieb and Shu, 1995b; Jackiewicz, 2003; Jackiewicz and Park, 2009; Keiner, 2009; Lurati, 2007; Phillips and Karageorghis, 1990; Vozovoi et al., 1996, 1997; Watanabe, 1990; Yilmazer and Kocar, 2008). To facilitate the presentation of the material that follows, we present in this section some useful background on the Gegenbauer polynomials using several results from approximation theory.

The Gegenbauer polynomial  $C_n^{(\alpha)}(x)$  of degree  $n \in \mathbb{Z}^+$  and associated with the parameter  $\alpha > -1/2$  is a real-valued function, and appears as an eigensolution to the following singular Sturm-Liouville problem in the finite domain  $[-1, 1]$  (Szegő, 1975):

$$\frac{d}{dx}(1-x^2)^{\alpha+\frac{1}{2}} \frac{dC_n^{(\alpha)}(x)}{dx} + n(n+2\alpha)(1-x^2)^{\alpha-\frac{1}{2}} C_n^{(\alpha)}(x) = 0. \quad (2.33)$$

The Gegenbauer polynomials  $C_n^{(\alpha)}(x)$ ,  $n = 0, 1, 2, \dots$  are defined as the coefficients in the following power series expansion of the generating function  $(1-2xt+t^2)^{-\alpha}$  (Horadam, 1985):

$$(1-2xt+t^2)^{-\alpha} = \sum_{n=0}^{\infty} C_n^{(\alpha)}(x)t^n, \quad |t| < 1.$$

They can also be generated through the following useful recursion formula:

$$(n+1)C_{n+1}^{(\alpha)}(x) = 2(n+\alpha)x C_n^{(\alpha)}(x) - (n+2\alpha-1)C_{n-1}^{(\alpha)}(x), \quad (2.34)$$

with the first two being  $C_0^{(\alpha)}(x) = 1$ ;  $C_1^{(\alpha)}(x) = 2\alpha x$ . The roots/zeros  $\{x_j\}_{j=0}^n$  of the Gegenbauer polynomial  $C_{n+1}^{(\alpha)}(x)$  are called the GG points, and the set

$$S_n^{(\alpha)} = \{x_j | C_{n+1}^{(\alpha)}(x_j) = 0, j = 0, \dots, n\}, \quad (2.35)$$

is called the set of GG points. The study of these GG points has been of quite interest because of their effect in many applications. For instance, the GG points of

## Chapter 2

---

the Gegenbauer polynomial  $C_n^{(\alpha)}(x)$  can be thought of as the positions of equilibrium of  $n \geq 2$  unit electrical charges in the interval  $(-1, 1)$  in the field generated by two identical charges of magnitude  $\alpha/2+1/4$  placed at 1 and  $-1$  (Ahmed et al., 1986). The weight function for the Gegenbauer polynomials is the even function  $w^{(\alpha)}(x) = (1 - x^2)^{\alpha-1/2}$ . The Gegenbauer polynomials form complete orthogonal basis polynomials in  $L^2_{w^{(\alpha)}}[-1, 1]$ . Their orthogonality relation is given by

$$\int_{-1}^1 w^{(\alpha)}(x) C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = h_n^{(\alpha)} \delta_{mn}, \quad (2.36)$$

where

$$h_n^{(\alpha)} = \frac{2^{1-2\alpha} \pi \Gamma(n + 2\alpha)}{n!(n + \alpha) \Gamma^2(\alpha)}, \quad (2.37)$$

is the normalization factor;  $\delta_{mn}$  is the Kronecker delta function. The symmetry of the Gegenbauer polynomials is emphasized by the relation (Hesthaven et al., 2007)

$$C_n^{(\alpha)}(x) = (-1)^n C_n^{(\alpha)}(-x). \quad (2.38)$$

Another suitable standardization of the Gegenbauer polynomials dates back to Doha (1990), where the Gegenbauer polynomials can be represented by

$$C_n^{(\alpha)}(x) = \frac{n! \Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} P_n^{(\alpha - \frac{1}{2}, \alpha - \frac{1}{2})}(x), \quad n = 0, 1, 2, \dots, \quad (2.39)$$

or equivalently

$$C_n^{(\alpha)}(1) = 1, \quad n = 0, 1, 2, \dots, \quad (2.40)$$

where  $P_n^{(\alpha - \frac{1}{2}, \alpha - \frac{1}{2})}(x)$  is the Jacobi polynomial of degree  $n$  and associated with the parameters  $\alpha - \frac{1}{2}; \alpha - \frac{1}{2}$ . This standardization establishes the useful relations that  $C_n^{(0)}(x)$  becomes identical with the Chebyshev polynomial of the first kind  $T_n(x)$ ,  $C_n^{(1/2)}(x)$  is the Legendre polynomial  $L_n(x)$ ;  $C_n^{(1)}(x)$  is equal to  $(1/(n+1))U_n(x)$ , where  $U_n(x)$  is the Chebyshev polynomial of the second type. Throughout the remaining of the dissertation, by the Gegenbauer polynomials we refer to those standardized by Equation (2.39) or Equation (2.40). Moreover, by the Chebyshev polynomials we refer to the Chebyshev polynomials of the first kind. Using the above standardization, the Gegenbauer polynomials are generated by Rodrigues' formula in the following form:

$$C_n^{(\alpha)}(x) = \left(-\frac{1}{2}\right)^n \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} (1 - x^2)^{\frac{1}{2} - \alpha} \frac{d^n}{dx^n} \left( (1 - x^2)^{n + \alpha - \frac{1}{2}} \right), \quad (2.41)$$

or starting with the following two equations:

$$C_0^{(\alpha)}(x) = 1, \quad (2.42a)$$

## Chapter 2

---

$$C_1^{(\alpha)}(x) = x, \quad (2.42b)$$

the Gegenbauer polynomials can be generated directly by the following three-term recurrence equation:

$$(j + 2\alpha)C_{j+1}^{(\alpha)}(x) = 2(j + \alpha)x C_j^{(\alpha)}(x) - j C_{j-1}^{(\alpha)}(x), \quad j \geq 1. \quad (2.42c)$$

Using Standardization (2.39) and Equation (4.7.1) in (Szegő, 1975), one can readily show that the Gegenbauer polynomials  $C_n^{(\alpha)}(x)$  and the Gegenbauer polynomials  $\hat{C}_n^{(\alpha)}(x)$  standardized by Szegő (1975) are related by

$$C_n^{(\alpha)}(x) = \frac{\hat{C}_n^{(\alpha)}(x)}{\hat{C}_n^{(\alpha)}(1)} \quad \forall x \in [-1, 1], \alpha > -\frac{1}{2}; n \geq 0. \quad (2.43)$$

Hence the Gegenbauer polynomials  $C_n^{(\alpha)}(x)$  satisfy the orthogonality relation

$$\int_{-1}^1 w^{(\alpha)}(x) C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = \lambda_n^{(\alpha)} \delta_{mn}, \quad (2.44)$$

where

$$\lambda_n^{(\alpha)} = \frac{2^{2\alpha-1} n! \Gamma^2(\alpha + \frac{1}{2})}{(n + \alpha) \Gamma(n + 2\alpha)}, \quad (2.45)$$

is the normalization factor. Moreover, the leading coefficients  $K_j^{(\alpha)}$  of the Gegenbauer polynomials  $C_j^{(\alpha)}(x)$  are

$$K_j^{(\alpha)} = 2^{j-1} \frac{\Gamma(j + \alpha) \Gamma(2\alpha + 1)}{\Gamma(j + 2\alpha) \Gamma(\alpha + 1)}, \quad (2.46)$$

for each  $j$ . The orthonormal Gegenbauer basis polynomials are defined by

$$\phi_j^{(\alpha)}(x) = (\lambda_j^{(\alpha)})^{-\frac{1}{2}} C_j^{(\alpha)}(x), \quad j = 0, \dots, n, \quad (2.47)$$

and they satisfy the following discrete orthonormality relation:

$$\sum_{j=0}^n \omega_j^{(\alpha)} \phi_s^{(\alpha)}(x_j) \phi_k^{(\alpha)}(x_j) = \delta_{sk}, \quad (2.48)$$

where

$$(\omega_j^{(\alpha)})^{-1} = \sum_{l=0}^n (\lambda_l^{(\alpha)})^{-1} (C_l^{(\alpha)}(x_j))^2; \quad x_j \in S_n^{(\alpha)}. \quad (2.49)$$

## Chapter 2

---

The integrations of the Gegenbauer polynomials  $C_j^{(\alpha)}(x)$  can be calculated exactly in terms of the Gegenbauer polynomials using Equations (2.42) as follows (El-Hawary et al., 2000):

$$\int_{-1}^x C_0^{(\alpha)}(x)dx = C_0^{(\alpha)}(x) + C_1^{(\alpha)}(x), \quad (2.50a)$$

$$\int_{-1}^x C_1^{(\alpha)}(x)dx = a_1(C_2^{(\alpha)}(x) - C_0^{(\alpha)}(x)), \quad (2.50b)$$

$$\int_{-1}^x C_j^{(\alpha)}(x)dx = \frac{1}{2(j+\alpha)}(a_2 C_{j+1}^{(\alpha)}(x) + a_3 C_{j-1}^{(\alpha)}(x) + (-1)^j(a_2 + a_3)), \quad j \geq 2, \quad (2.50c)$$

where

$$a_1 = \frac{1+2\alpha}{4(1+\alpha)}, \quad a_2 = \frac{j+2\alpha}{(j+1)}, \quad a_3 = -\frac{j}{(j+2\alpha-1)}.$$

For further information about the Gegenbauer polynomials, the interested reader may consult (Abramowitz and Stegun, 1965; Bayin, 2006; Szegö, 1975).

## 2.4 The Gegenbauer Approximation of Functions

The approximation of a function by a truncated series of basis functions is the fundamental idea in spectral methods, and the choice of the expansion basis functions largely influences the superior approximation properties of spectral methods relative to other methods such as the finite difference schemes and the finite element methods. The expansion functions must conveniently have three basic properties: (i) Ease of evaluation, (ii) completeness; (iii) orthogonality. Property (i) is quite essential, and is the main reason behind the application of the trigonometric functions and polynomials in the discretization process. Property (ii) is also necessary so that each function of a given space can be conveniently represented as a limit of a linear combination of such basis functions. Property (iii) is extremely important to establish the fast conversion between the coefficients of the spectral expansion series and the values of the function at some certain nodes  $\{x_i\}_{i=0}^N$  (Fornberg, 1996). This last property is the key for the study of many properties of the classical orthogonal basis polynomials and their intensive applications.

The Gegenbauer polynomials satisfy all of the above three properties. Indeed, they can be directly generated through the three-term recurrence relations (2.42). They form a complete orthogonal basis system in  $L^2([-1, 1], w^{(\alpha)}(x))$ , so that a function  $y(x) \in C^0[-1, 1]$  can be expanded as an infinite series of the infinitely differentiable global Gegenbauer basis polynomials. In particular, a natural approximation in a classical Gegenbauer expansion method is sought in the form of

## Chapter 2

---

the truncated series

$$y(x) \approx \sum_{k=0}^N a_k C_k^{(\alpha)}(x), \quad (2.51)$$

where  $a_k$  are the Gegenbauer spectral expansion coefficients of the solution typically determined by variational principles or by the weighted-residual methods (Villadsen and Stewart, 1995). The truncation error produced by the Gegenbauer expansion series (2.51) for a smooth function  $y(x)$  defined on  $[-1, 1]$  is given by the following theorem:

**Theorem 2.4.1** (Truncation error (El-Hawary et al., 2000)). *Let  $y(x) \in C^\infty[-1, 1]$  be approximated by the Gegenbauer expansion series (2.51), then for each  $x \in [-1, 1]$ , a number  $\xi(x) \in [-1, 1]$  exists such that the truncation error  $E_T(x, \xi, N, \alpha)$  is given by*

$$E_T(x, \xi, N, \alpha) = \frac{y^{(N+1)}(\xi)}{(N+1)!K_{N+1}^{(\alpha)}} C_{N+1}^{(\alpha)}(x). \quad (2.52)$$

Theorem 2.4.1 shows that the error term is the monic polynomial  $\phi_N^{(\alpha)} = (y^{(N+1)}(\xi)C_{N+1}^{(\alpha)}(x))/((N+1)!K_{N+1}^{(\alpha)})$ , which can be derived from the standard Cauchy remainder term in the error formula of polynomial interpolation. The following section highlights the convergence rate of the Gegenbauer collocation methods.

## 2.5 Gegenbauer Collocation Methods: Convergence Rate

In a Gegenbauer collocation method, the Gegenbauer collocation coefficients  $a_k$  are evaluated by requiring that the approximation must match the function values  $y(x_i)$  for a certain collocation points set  $\{x_i\}_{i=0}^N$ , i.e. to obtain the discrete values of the Gegenbauer coefficients, the following interpolation conditions are imposed:

$$y(x_i) = \sum_{k=0}^N a_k C_k^{(\alpha)}(x_i), \quad i = 0, \dots, N. \quad (2.53)$$

Since the zeros of the appropriate orthogonal polynomials yield better accuracy than the uniformly distributed collocation points (Gottlieb and Hesthaven, 2001; Oh and Luus, 1977), the  $(N+1)$  collocation points are frequently chosen to be the interior GG points  $x_i \in S_N^{(\alpha)}$  as they satisfy the attractive density distribution (2.32), and cluster quadratically near the endpoints of the solution domain  $[-1, 1]$ . In this case, the convergence rate of the Gegenbauer expansion series (2.51) can

## Chapter 2

---

be described by the following theorem, which dates back to the seminal work of Gottlieb and Shu (1995a):

**Theorem 2.5.1** (Gegenbauer collocation convergence rate (Gottlieb and Shu, 1995a)). *Let*

$$\tilde{y}_N(x) = \sum_{k=0}^N a_k C_k^{(\alpha)}(x), \quad (2.54)$$

*be the Gegenbauer collocation approximation of the function  $y(x)$  with the weight  $w^{(\alpha)}(x) = (1 - x^2)^{\alpha-1/2}$ , where the Gegenbauer collocation coefficients are computed by interpolating the function  $y(x)$  at the GG points. If the function  $y(x) \in C^K[-1, 1]$ , then the Gegenbauer collocation approximation converges exponentially in the sense that*

$$\|y - \tilde{y}\|_{L^2_{w^{(\alpha)}}} \leq \frac{A}{N^K} \|y^{(K)}\|_{L^\infty}, \quad (2.55)$$

*where the weighted  $L^2$ -norm is defined by*

$$\|y\|_{L^2_{w^{(\alpha)}}}^2 = \int_{-1}^1 w^{(\alpha)}(x) |y(x)|^2 dx, \quad (2.56)$$

*and  $A$  is a constant independent of  $N$  and  $K$ .*

Theorem 2.5.1 shows that the rate of convergence of the error to zero is contingent on the regularity of the function  $y(x)$ . A discontinuity in the solution function  $y(x)$ , or in any of its derivatives results in a reduced order of convergence. Typically, for a solution function  $y(x) \in C^K[-1, 1]$ ,  $K \in \mathbb{Z}^+$ , the truncated series (2.51) converges algebraically with  $O(N^{-K})$ -convergence (Gottlieb et al., 2011). For analytic functions, the produced approximation error of the weighted sum of the smooth Gegenbauer basis polynomials (2.51) approaches zero with an exponential rate of  $O(N^{-N})$  when  $N$  tends to infinity, which is faster than any polynomial rate for smooth functions. This rapid convergence characteristic of the Gegenbauer expansion methods is broadly-known as the spectral accuracy. In contrast, the global error for a finite-difference method with  $N$  grid points scales as  $N^{-p}$ , where  $p$  is the fixed order of the method (Barranco and Marcus, 2006). The fast convergence property of the Gegenbauer expansion method, as inherited from spectral methods, compared to finite difference schemes is largely due to the global nature of the Gegenbauer approximation in the sense that a computation at any point in the solution domain depends on the information from the whole domain of computation, not only on information at neighboring points. Hence the chief advantage of the Gegenbauer approximation methods over other classical approximation schemes such as the finite-difference methods lies in the achieved

## Chapter 2

---

accuracy per degrees of freedom (i.e. the number of Gegenbauer modes, or the number of collocation points), where it is possible to attain very good accuracy with relatively coarse grids. Indeed, as we shall demonstrate later in Chapters 5 and 7, to get the same level of accuracy using classical discretization methods, a Gegenbauer collocation method generally requires far fewer degrees of freedom.

### 2.6 The Gegenbauer Operational Matrix of Integration

The concept of the operational matrix of integration was originally invented by the Egyptian scientist El-Gendi in the year 1969. El-Gendi (1969) noticed that the definite integrals  $\int_{-1}^{x_i} f_N(x)dx$  of a truncated Chebyshev expansion series  $f_N(x)$  approximating a well-behaved function  $f(x)$  in  $[-1, 1]$  can be easily represented by a square matrix, for a certain set of grid points  $\{x_i\}_{i=0}^N$ . The idea emerged after Clenshaw and Curtis (1960) presented their popular procedure for the numerical integration of a continuous and of bounded variation function  $f(x)$  defined on a finite range  $-1 \leq x \leq 1$ , by expanding the spectral interpolant  $f_N(x)$  in a truncated Chebyshev polynomials series as follows:

$$f_N(x) = \sum_{k=0}^N {}'' a_k T_k(x), \quad (2.57)$$

where

$$a_k = \frac{2}{N} \sum_{j=0}^N {}'' f(x_j) T_k(x_j), \quad (2.58)$$

$$x_j = \cos\left(\frac{j\pi}{N}\right), \quad j = 0, 1, \dots, N, \quad (2.59)$$

and the summation symbol with double primes denotes a sum with both the first and last terms halved. Here the points  $x_j, j = 0, 1, \dots, N$ , are the Chebyshev-Gauss-Lobatto (CGL) grid points. The indefinite integral of the function  $f(x)$  can be approximated in the spectral space by integrating the truncated Chebyshev expansion series (2.57) term by term.

El-Gendi (1969) proposed to approximate the definite integrals of the function  $f(x)$  in the physical space instead by multiplying a constant matrix with the vector  $F = (f(x_0), f(x_1), \dots, f(x_N))^T$  of the function values calculated at the CGL points. This idea can be carried out by first expressing the indefinite integrals in terms of the Chebyshev polynomials themselves as follows:

$$\int_{-1}^x f_N(t) dt = \sum_{j=0}^N {}'' a_j \int_{-1}^x T_j(t) dt = \sum_{j=0}^{N+1} \hat{c}_j T_j(x), \quad (2.60)$$

## Chapter 2

---

where

$$\hat{c}_0 = \sum_{j=0, j \neq 1}^N \frac{(-1)^{j+1} a_j}{j^2 - 1} - \frac{1}{4} a_1, \quad (2.61a)$$

$$\hat{c}_k = \frac{a_{k-1} - a_{k+1}}{2k}, \quad k = 1, 2, \dots, N-2, \quad (2.61b)$$

$$\hat{c}_{N-1} = \frac{a_{N-2} - 0.5a_N}{2(N-1)}, \quad (2.61c)$$

$$\hat{c}_N = \frac{a_{N-1}}{2N}; \quad (2.61d)$$

$$\hat{c}_{N+1} = \frac{a_N}{4(N+1)}. \quad (2.61e)$$

Through Equations (2.58) and (2.61), one can derive the following equation:

$$\left( \int_{-1}^{x_0} f_N(x) dx, \int_{-1}^{x_1} f_N(x) dx, \dots, \int_{-1}^{x_N} f_N(x) dx \right)^T = B^{(1)} F, \quad (2.62)$$

where  $B^{(1)} = (b_{i,j}^{(1)})$  is the first-order Chebyshev square integration matrix of size  $(N+1)$ ;  $b_{i,j}^{(1)}, 0 \leq i, j \leq N$ , are the elements of the integration matrix  $B^{(1)}$ . Hence, in general, an integration matrix is simply a linear map which takes a vector of  $N$  function values  $f(x_i)$  to a vector of  $N$  integral values  $\int_a^{x_i} f(x) dx$ , for some real number  $a \in \mathbb{R}$ . The introduction of the numerical integration matrix has provided the key to apply the rich and powerful matrix linear algebra in many areas (Babolian and Fattahzadeh, 2007; Danfu and Xufeng, 2007; Elgindy, 2009; Elgindy and Hedar, 2008; Endow, 1989; Guf and Jiang, 1996; Paraskevopoulos et al., 1985; Razzaghi et al., 1990; Razzaghi and Yousefi, 2001; Williams, 2006; Wu, 2009).

Similarly, in a Gegenbauer collocation method based on the Gauss points  $x_i \in S_N^{(\alpha)}$ , one can define the definite integrals of the Gegenbauer collocation approximation  $\tilde{y}_N(x)$  of the function  $y(x)$  through a matrix-vector multiplication in the following form:

$$\left( \int_{-1}^{x_0} \tilde{y}_N(x) dx, \int_{-1}^{x_1} \tilde{y}_N(x) dx, \dots, \int_{-1}^{x_N} \tilde{y}_N(x) dx \right)^T = Q^{(1)} Y, \quad (2.63)$$

where  $Q^{(1)} = (q_{i,j}^{(1)}), 0 \leq i, j \leq N$ , is the first-order GIM;  $Y = (\tilde{y}_N(x_0), \tilde{y}_N(x_1), \dots, \tilde{y}_N(x_N))^T$ .

The GIM was first developed by El-Hawary et al. in the year 2000. Their approach for constructing the elements of the GIM was originally outlined in the following theorem:

## Chapter 2

---

**Theorem 2.6.1** ((El-Hawary et al., 2000)). *Let  $f(x)$  be approximated by the Gegenbauer polynomials;  $x_k \in S_N^{(\alpha)}$ , then there exist a matrix  $Q^{(1)} = (q_{ij}^{(1)})$ ,  $i, j = 0, \dots, N$ ; and a number  $\xi = \xi(x) \in [-1, 1]$  satisfying*

$$\int_{-1}^{x_i} f(x) dx = \sum_{k=0}^N q_{ik}^{(1)}(\alpha) f(x_k) + E_N^{(\alpha)}(x_i, \xi), \quad (2.64a)$$

where

$$q_{ik}^{(1)}(\alpha) = \sum_{j=0}^N (\tilde{\lambda}_j^{(\alpha)})^{-1} \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k) \int_{-1}^{x_i} C_j^{(\alpha)}(x) dx, \quad (2.64b)$$

$$(\omega_k^{(\alpha)})^{-1} = \sum_{j=0}^N (\tilde{\lambda}_j^{(\alpha)})^{-1} (C_j^{(\alpha)}(x_k))^2, \quad (2.64c)$$

$$\tilde{\lambda}_j^{(\alpha)} = 2^{j+2\alpha+\tau} j! \frac{\Gamma(\alpha + \frac{1}{2}) \Gamma(j + \alpha + \frac{1}{2})}{\Gamma(2j + 2\alpha + 1)} \tilde{K}_j^{(\alpha)}, \quad (2.64d)$$

$$\tau = \begin{cases} 1, & \text{if } \alpha = j = 0, \\ 0, & \text{otherwise,} \end{cases} \quad (2.64e)$$

$$\tilde{K}_j^{(\alpha)} = 2^j \frac{\Gamma(j + \alpha) \Gamma(2\alpha + 1)}{\Gamma(j + 2\alpha) \Gamma(\alpha + 1)}, \quad (2.64f)$$

$$E_N^{(\alpha)}(x, \xi) = \frac{f^{(N+1)}(\xi)}{(N+1)! \tilde{K}_{N+1}^{(\alpha)}} \int_{-1}^x C_{N+1}^{(\alpha)}(x) dx. \quad (2.64g)$$

Equation (2.64a) provides the Gegenbauer quadrature approximation, Equation (2.64b) defines the elements of the first-order Q-matrix,  $Q^{(1)}$ , calculated at  $S_N^{(\alpha)}$ . Equations (2.64c)–(2.64f) define the required parameters for the construction of  $Q^{(1)}$ ; Equation (2.64g) defines the error term. To achieve the best possible approximation using the Q-matrix, El-Hawary et al. (2000) further provided a means to optimize the selection procedure of the Gegenbauer parameter  $\alpha$  under a certain optimality measure. In the next chapter, we shall explore this optimality measure, and discover that their presented numerical scheme is associated with many drawbacks which limit its application in practice. We shall also provide a strong and practical numerical method for the construction of an optimal Gegenbauer quadrature built upon the strengths of the popular Chebyshev, Legendre, and Gegenbauer polynomials.

### 2.7 Solving Various Dynamical Systems and OC Problems by Optimizing the GIM

The Gegenbauer polynomials are very versatile and have been applied extensively in many research areas such as studying annihilation processes, improving tissue segmentation of human brain magnetic resonance imaging, queuing theory, resolution of the Gibbs phenomenon by reconstructing piecewise smooth functions in smooth intervals with exponential accuracy up to the edges of the interval, analysis of light scattering from homogeneous dielectric spheres, calculation of complicated Feynman integrals, numerical quadratures, flutter analysis of an airfoil with bounded random parameters in incompressible flow, studying third-order nonlinear systems, solving ODEs and PDEs, solving OC problems, etc.; cf. (Archibald et al., 2003; Bavinck et al., 1993; Ben-yu, 1998; Doha and Abd-Elhameed, 2002; Doha, 1990; El-Hawary et al., 2000, 2003; Gelb, 2004; Gelb and Gottlieb, 2007; Gottlieb and Shu, 1995b; Kotikov, 2001; Lampe and Kramer, 1983; Ludlow and Everitt, 1995; Srirangarajan et al., 1975; Vozovoi et al., 1996; Wu et al., 2007). In a standard Gegenbauer polynomial approximation method, the unknown solution is expanded by the Gegenbauer expansion series (2.51) using a predefined Gegenbauer parameter value  $\alpha$ . To achieve better solution approximations, some methods presented in the literature apply the GIM for approximating the integral operations, and recast various mathematical problems such as ODEs, integral and integro-differential equations, and OC problems into unconstrained/constrained optimization problems. The Gegenbauer parameter  $\alpha$  associated with the Gegenbauer polynomials is then added as an extra unknown variable to be optimized in the resulting optimization problem as an attempt to optimize its value rather than choosing a random value. Although this idea of optimizing the GIM to gain more accuracy in the approximations is tempting, later in Chapter 4 we shall prove theoretically that this optimization procedure is not possible as it violates the discrete Gegenbauer orthonormality relation, and may in turn produce false solution approximations. In Chapters 5-7, we shall discover that more practical and robust Gegenbauer collocation schemes can be established using the GIMs developed in the next chapter.

## PART B: Suggested Declaration for Thesis Chapter

Monash University

### Declaration for Thesis Chapter 3

#### Declaration by candidate

In the case of Chapter 3, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
The author of the key ideas, programming codes, organization, development, and writing up of the article	90

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided valuable comments and aided proofreading	

Candidate's Signature

	Date 11/05/2013
--	--------------------

#### Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)

Signature 1

	Date 14/5/13
---	--------------

.....

This page is intentionally left blank

## *Chapter 3*

# *Optimal Gegenbauer Quadrature Over Arbitrary Integration Nodes*

**Chapter 3 is based on the published article Elgindy, K. T., Smith-Miles, K. A., April 2013. Optimal Gegenbauer quadrature over arbitrary integration nodes. Journal of Computational and Applied Mathematics 242 (0), 82–106.**

**Abstract.** *This chapter treats definite integrations numerically using Gegenbauer quadratures. The novel numerical scheme introduces the idea of exploiting the strengths of the Chebyshev, Legendre, and Gegenbauer polynomials through a unified approach, and using a unique numerical quadrature. In particular, the numerical scheme developed employs the Gegenbauer polynomials to achieve rapid rates of convergence of the quadrature for the small range of the spectral expansion terms. For a large-scale number of expansion terms, the numerical quadrature has the advantage of converging to the optimal Chebyshev and Legendre quadratures in the  $L^\infty$ -norm and  $L^2$ -norm, respectively. The key idea is to construct the Gegenbauer quadrature through discretizations at some optimal sets of points of the Gegenbauer-Gauss (GG) type in a certain optimality sense. We show that the Gegenbauer polynomial expansions can produce higher-order approximations to the definite integrals  $\int_{-1}^{x_i} f(x)dx$  of a smooth function  $f(x) \in C^\infty[-1, 1]$  for the small range by minimizing the quadrature error at each integration point  $x_i$  through a pointwise approach. The developed Gegenbauer quadrature can be applied for approximating integrals with any arbitrary sets of integration nodes. Exact integrations are obtained for polynomials of any arbitrary degree  $n$  if the number of columns in the developed Gegenbauer integration matrix (GIM) is greater than or equal to  $n$ . The error formula for the Gegenbauer quadrature is derived. Moreover, a study on the error bounds and the convergence rate shows that the optimal Gegenbauer quadrature exhibits very rapid convergence rates faster than any finite power of the number of Gegenbauer expansion terms. Two efficient computational algorithms are presented for optimally constructing the Gegenbauer quadrature. We illustrate the high-order approximations of the optimal Gegenbauer quadrature through extensive numerical experiments including comparisons with conventional Chebyshev, Legendre, and Gegenbauer polynomial expansion methods. The present method is broadly applicable and represents a strong addition to the arsenal of numerical quadrature methods.*

**Keyword.** *Gegenbauer-Gauss points; Gegenbauer integration matrix; Gegenbauer polynomials; Gegenbauer quadrature; Numerical integration; Spectral methods.*

References are considered at the end of the thesis.

## Chapter 3

# Optimal Gegenbauer Quadrature Over Arbitrary Integration Nodes

### 3.1 Introduction

Numerical integrations have found numerous applications in many scientific areas; cf. (El-Gendi, 1969; Elbarbary, 2006, 2007; Elgindy, 2009; Elgindy and Smith-Miles, 2013c; Elgindy et al., 2012; Ghoreishi and Hosseini, 2008; Greengard, 1991; Lee and Tsay, 1989; Mai-Duy and Tanner, 2007; Marzban and Razzaghi, 2003; Mihaila and Mihaila, 2002; Paraskevopoulos, 1983; Tian, 1989). In particular, they frequently arise in the solution of ordinary differential equations, partial differential equations, integral equations, integro-differential equations, optimal control problems, etc. The increasing range and significance of their applications manifest the demand for achieving higher-order quadrature approximations using robust and efficient numerical algorithms. The most straightforward numerical integration technique uses the Newton-Cotes formulas; however, Gaussian quadratures are known to produce the most accurate approximations possible through choosing the zeros of the orthogonal polynomials and their corresponding weighting functions (Weisstein, 2003). Among the classical orthogonal polynomials commonly used are the Jacobi polynomials, which appear as eigenfunctions of singular Sturm-Liouville problems. Their applications give rise to the elegant class of methods known as the spectral methods. The growing interest in these methods is largely due to their promise of “spectral accuracy” if the function being represented is infinitely smooth, and their superior approximation properties compared with other methods of discretization (Gottlieb and Orszag, 1977). In particular, for sufficiently smooth functions, the  $k$ th coefficient of the spectral

## Chapter 3

---

expansion decays faster than any inverse power of  $k$ . Consequently very good approximations to the function are obtained with relatively few terms (Breuer and Everson, 1992).

In a classical spectral method, the function  $f(x) \in C^\infty[-1, 1]$  is expanded in terms of trial functions  $\{\phi_k(x)\}_{k=0}^n$  as a finite series of the form  $f(x) \approx \sum_{k=0}^n \hat{f}_k \phi_k(x)$ , where  $\{\hat{f}_k\}_{k=0}^n$  are the spectral coefficients. The trial functions are globally smooth functions, and their choice usually depends on the type of the underlying problem. It is widely known that the trigonometric polynomials (Fourier series) are favored for periodic problems, while Jacobi polynomials are considered excellent basis polynomials for non-periodic problems (Fornberg, 1996). Jacobi polynomials include the Gegenbauer (ultraspherical) polynomials  $C_n^{(\alpha)}(x)$  (see Appendix 3.A), the Chebyshev polynomials  $T_n(x)$ , and Legendre polynomials  $L_n(x)$ . The latter two are special cases of the Gegenbauer polynomials for the Gegenbauer parameter values  $\alpha = 0; 0.5$ , respectively (Boyd, 2006). For decades, Chebyshev and Legendre polynomials have attracted much attention due to their fast convergence properties. However, we find some special results and clear reasons in the literature which motivate us to apply a unified approach using the Gegenbauer polynomials rather than applying the standard Chebyshev and Legendre polynomial approximations. For instance, (i) it is well-known that expansions in Chebyshev polynomials are better suited to the solution of hydrodynamic stability problems than expansions in other sets of orthogonal functions (Orszag, 1971). On the other hand, in the resolution of thin boundary layer applications, Legendre polynomial expansions give exceedingly good representations of functions that undergo rapid changes in narrow boundary layers (Gottlieb and Orszag, 1977). Hence, it is convenient to apply a unified approach using the Gegenbauer polynomials, which include the Legendre and the Chebyshev polynomials as special cases, to capture the most suitable property requirements for a given problem, rather than applying the particular choices of the Chebyshev and Legendre polynomials for various approximation problems. Moreover, the theoretical and experimental results derived in a Gegenbauer polynomial approximation method apply directly to Chebyshev and Legendre approximation methods as special cases. (ii) Light's work (Light, 1978) on the computed norms of some Gegenbauer projection operators confirms that the Chebyshev and Legendre projections cannot be minimal as they all increase monotonically with  $\alpha$  (Mason and Handscomb, 2003). In particular, the reported results show that the norm of the Chebyshev projection is not the smallest for Chebyshev series expansions truncated after  $n$  terms in the range  $1 < n < 10$ . (iii) The work of Doha (1990) in approximating the solution of boundary value problems (BVPs) for linear partial differential equations in one dimension shows that higher-order approximations better than those obtained from Chebyshev and Legendre poly-

## Chapter 3

---

nomials can be obtained from Gegenbauer polynomial expansions for small and negative values of  $\alpha$ . (iv) The work of El-Hawary et al. (2000) on the numerical approximation of definite integrations using Gegenbauer integration matrices (GIMs) shows an advantage of the Gegenbauer polynomials over the Chebyshev and Legendre polynomials. (v) The recent works of Elgindy and Smith-Miles (2013c) and Elgindy et al. (2012) show that the Gegenbauer polynomial methods are very effective in the solutions of BVPs, integral equations, integro-differential equations; and optimal control problems. Moreover, the reported results illustrate that some members of the Gegenbauer family of polynomials converge to the solutions of the problems faster than Chebyshev and Legendre polynomials for the small/medium range of the number of spectral expansion terms.

The Gegenbauer polynomials have already been applied extensively in many research areas, and have been demonstrated to provide excellent approximations of analytic functions; cf. (Archibald et al., 2003; Barrio, 1999; Doha and Abd-Elhameed, 2009; Gelb, 2004; Gottlieb and Shu, 1995b; Lurati, 2007; Malek and Phillips, 1995; Phillips and Karageorghis, 1990; Vozovoi et al., 1996, 1997; Yilmazer and Kocar, 2008). The present work in this chapter introduces a strong and practical numerical method for the construction of optimal GIMs to efficiently approximate definite integrations. The proposed quadrature method outperforms the numerical method presented earlier by El-Hawary et al. (2000), which suffers from several major drawbacks raised in Section 3.2. The significant contribution of this chapter is in the introduction of a novel Gegenbauer quadrature method which takes advantage of the major strengths of the three orthogonal polynomials, namely the Chebyshev, Legendre, and Gegenbauer polynomials. In particular, the novel quadrature exploits the rapid convergence properties of the Gegenbauer polynomials for the small/medium range of the number of spectral expansion terms, and converges to the optimal Chebyshev quadrature in the  $L^\infty$ -norm for a large-scale number of expansion terms. The proposed quadrature can also be manipulated easily to converge to the Legendre quadrature in the  $L^2$ -norm, for a large-scale number of expansion terms. The proposed method treats the definite integrals  $\int_{-1}^{x_i} f(x)dx$ , for some given function  $f(x)$ , separately for each integration point  $x_i$  using distinct optimal sets of interpolation/discretization points in the sense of solving Problem (3.13); cf. Section 3.2.1. We show that faster convergence rates of the Gegenbauer expansion series can be achieved for the small/medium range of the expansion terms using some optimal values of the Gegenbauer parameter  $\alpha$ , which damp the quadrature error at each integration point through a pointwise approach. The framework for constructing the Gegenbauer quadrature is based on the integration points set, regardless of the integrand function. Moreover, the proposed technique allows for the approximation of definite integrals for any arbitrary sets of integration points. The efficiency of the proposed numerical quadrature increases for symmetric sets of integration

## Chapter 3

---

points, where most of the calculations in the algorithms developed are halved; cf. Section 3.2.9. The extensive numerical experiments conducted in Section 3.3 show an advantage of the optimal Gegenbauer quadrature over the standard Chebyshev, Legendre, and Gegenbauer quadratures.

The remainder of this chapter is organized as follows: In the following section we introduce the GIMs, and the El-Hawary et al. (2000) approach for optimizing them in the sense of solving Problem (3.8) to produce higher-order approximations to the definite integrals. Moreover, we highlight some of the main issues associated with their method. In section 3.2.1, we propose our optimal Gegenbauer quadrature method for approximating definite integrals in the sense of solving Problem (3.13). We describe the procedure for constructing the novel optimal GIM, and derive an error formula for the truncation error of the Gegenbauer quadrature in Section 3.2.2. We briefly highlight the error in the polynomial integration in Section 3.2.3. In Section 3.2.4, we study the bounds on the Gegenbauer quadrature error and the convergence rate. In Section 3.2.5, we prove the convergence of the Gegenbauer quadrature to the Chebyshev quadrature in the  $L^\infty$ -norm. In Section 3.2.6, we determine the suitable interval of uncertainty for the optimal Gegenbauer parameters of the GIM for small/medium range expansions. In Section 3.2.7, we highlight some substantial advantages of the Gegenbauer collocation methods endowed with the optimal GIM for the discretization of various continuous mathematical models. In Section 3.2.8, we establish the Gegenbauer approximations of definite integrals in matrix form. In Section 3.2.9, we develop two efficient computational algorithms for ideally constructing the optimal GIM. In Section 3.3, we report some extensive numerical results demonstrating the efficiency and accuracy of our proposed Gegenbauer quadrature method through comparisons with standard Chebyshev and Gegenbauer quadratures. Some further applications of the proposed Gegenbauer quadrature method are highlighted in Section 3.4. Section 3.5 opens the door for future directions on the proposed Gegenbauer quadrature, and is followed by some concluding remarks in Section 3.6. We briefly present some useful properties of the Gegenbauer polynomials in 3.A. The computational algorithms and the proofs of various theorems and lemmas are included in Appendices 3.B–3.H.

### 3.2 Generation of Optimal GIMs

In a typical spectral method approximating a function  $f(x) \in C^\infty[-1, 1]$  using Gegenbauer polynomials, the function  $f(x)$  is approximated by a truncated

## Chapter 3

---

Gegenbauer expansion series as follows:

$$f(x) \approx \sum_{k=0}^N a_k C_k^{(\alpha)}(x), \quad (3.1)$$

where  $a_k$  are the Gegenbauer coefficients. The integration of the function  $f(x)$  is approximated by integrating the finite Gegenbauer expansion series, and the sought definite integration approximations for a certain set of integration nodes  $\{x_i\}_{i=0}^N$  can be expressed in a matrix-vector multiplication form as follows:

$$I = \begin{bmatrix} \int_{-1}^{x_0} f(x) dx \\ \int_{-1}^{x_1} f(x) dx \\ \vdots \\ \int_{-1}^{x_N} f(x) dx \end{bmatrix} = \begin{bmatrix} \hat{p}_{00} & \cdots & \hat{p}_{0N} \\ \vdots & \ddots & \vdots \\ \hat{p}_{N0} & \cdots & \hat{p}_{NN} \end{bmatrix} \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix} = \hat{P}F, \quad (3.2)$$

where the matrix  $\hat{P} = (\hat{p}_{i,j}), 0 \leq i, j \leq N$ , is the Gegenbauer operational matrix of integration, and is usually referred to as the GIM. Using Equations (3.A.7) & (3.A.8), and following the method presented by El-Hawary et al. (2000), one can readily construct the elements of the  $\hat{P}$ -matrix through the following theorem:

**Theorem 3.2.1.** *Let*

$$S_N^{(\alpha)} = \{x_k | C_{N+1}^{(\alpha)}(x_k) = 0, k = 0, \dots, N\}, \quad (3.3)$$

*be the set of Gegenbauer-Gauss (GG) points. Moreover, let  $f(x) \in C^\infty[-1, 1]$  be approximated by the Gegenbauer expansion series (3.1); then there exist a matrix  $\hat{P} = (\hat{p}_{i,j}), 0 \leq i, j \leq N$ ; and some numbers  $\xi_i \in [-1, 1]$  satisfying*

$$\int_{-1}^{x_i} f(x) dx = \sum_{k=0}^N \hat{p}_{ik}(\alpha) f(x_k) + E_N^{(\alpha)}(x_i, \xi_i), \quad (3.4)$$

where

$$\hat{p}_{ik}(\alpha) = \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k) \int_{-1}^{x_i} C_j^{(\alpha)}(x) dx, \quad (3.5)$$

$$(\omega_k^{(\alpha)})^{-1} = \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} (C_j^{(\alpha)}(x_k))^2, \quad x_k \in S_N^{(\alpha)}, \quad (3.6)$$

$$E_N^{(\alpha)}(x_i, \xi_i) = \frac{f^{(N+1)}(\xi_i)}{(N+1)! K_{N+1}^{(\alpha)}} \int_{-1}^{x_i} C_{N+1}^{(\alpha)}(x) dx, \quad (3.7)$$

and  $\lambda_j^{(\alpha)}; K_{N+1}^{(\alpha)}$  are as defined by Equations (3.A.7) & (3.A.8), respectively.

## Chapter 3

---

The  $\hat{P}$ -matrix is a square GIM with a fixed  $\alpha > -1/2$  value, which is identical to the Chebyshev and Legendre matrices on setting  $\alpha = 0; 0.5$ , respectively. Here  $E_N^{(\alpha)}(x_i, \xi_i)$  represents the error term produced by the  $\hat{P}$ -matrix quadrature for each  $x_i$ . This error term is the integral of the monic polynomial  $(f^{(N+1)}(\xi_i)C_{N+1}^{(\alpha)}(x))/((N+1)!K_{N+1}^{(\alpha)})$ , which can be derived from the standard Cauchy remainder term in the error formula of polynomial interpolation. Since the error term depends on the Gegenbauer parameter  $\alpha$ , a natural idea for damping the quadrature error is to optimally control the value of  $\alpha$  in the  $\hat{P}$ -matrix quadrature (3.4) rather than choosing any random  $\alpha$  value. This approach produces an optimal GIM in the sense of minimizing the Gegenbauer quadrature truncation error, and avoids the degradation of precision produced by choosing random  $\alpha$  values. Perhaps one of the earliest methods for optimizing the value of  $\alpha$  and constructing an optimal GIM was that proposed by El-Hawary et al. in 2000 through the solution of the following constrained optimization problem:

$$\text{Find } \alpha = \alpha^* \text{ which minimizes } J = \left( \int_{-1}^1 \left| E_N^{(\alpha)}(x, \xi) \right|^p d\xi \right)^{1/p}, \quad \alpha > -1/2; p \rightarrow \infty. \quad (3.8)$$

This constrained optimization problem can be transformed into an unconstrained optimization problem through the change of variable

$$\alpha = e^{(t^2 + \varepsilon)} - \frac{3}{2}, \quad 0 < \varepsilon \ll 1. \quad (3.9)$$

The numerical quadrature then approximates the definite integrals of the function  $f(x)$  by interpolating the function at the GG points  $S_N^{(\alpha^*)}$ , and integrating the Gegenbauer interpolant term by term. In fact, this method successfully achieves higher-order approximations which exceed the precision of the Chebyshev and Legendre polynomial expansion methods on some test problems as shown in (El-Hawary et al., 2000). However, there are several drawbacks associated with this numerical scheme: (i) For increasing values of the exponent  $p$ , the values of  $\left| E_N^{(\alpha)}(x, \xi) \right|^p$  may grow so large that they become computationally prohibitive. (ii) Solving the one-dimensional optimization problem (3.8) entails the evaluation of the  $(N+1)$ th derivative of  $f$  at each iteration for increasing values of  $p$  until the reduced optimization problem approximates the minimax problem:

$$\text{Find } \alpha^* = \operatorname{argmin}_{\alpha > -1/2} \max_{-1 < \xi < 1} \left| E_N^{(\alpha)}(x, \xi) \right|. \quad (3.10)$$

This reduces the efficiency of the quadrature method if the function  $f$  is so complicated that the evaluations of its derivatives are very expensive and time-consuming. (iii) The numerical approximation of the  $(N+1)$ th derivative of  $f$  is

## Chapter 3

---

prone to large round-off errors for increasing values of  $N$ , since numerical differentiation is in principle an ill-posed problem (Liu et al., 2011). (iv) The method does not provide useful means for performing numerical integrations for general arbitrary sets of integration nodes. On the contrary, the method seems to work well only at the GG points  $x_i \in S_N^{(\alpha^*)}$ . (v) Higher-order approximations cannot be achieved unless the number of integration nodes ( $N + 1$ ) is increased, since the  $\hat{P}$ -matrix is a square matrix of size  $(N + 1)$ . (vi) The method of construction of the Gegenbauer quadrature is principally contingent on the form of the integrand function  $f(x)$ , which prevents the automatic construction of the numerical quadrature. All of these problems show the need for developing a new approach able to sustain higher-order approximations for general sets of integration nodes, and avoids the aforementioned problems.

**Remark 3.2.2.** *The  $\hat{P}$ -matrix presented in this section is a modified version of the  $Q$ -matrix, which was originally outlined by El-Hawary et al. (2000). In the remainder of this article, when we refer to the  $\hat{P}$ -matrix or  $Q$ -matrix we shall mean the standard  $\hat{P}$ -matrix or  $Q$ -matrix constructed using any arbitrary choice of the Gegenbauer parameter  $\alpha$ ; the acronyms  $\hat{P}MQ$  and  $QMQ$  refer to their associated quadratures, respectively. Moreover, in referring to the optimal  $Q$ -matrix, we shall mean the  $Q$ -matrix constructed by solving Problem (3.8).*

### 3.2.1 The Proposed Method

We propose to construct an optimal Gegenbauer quadrature by minimizing the magnitude of the quadrature error  $E_N^{(\alpha)}(x, \xi)$  at each integration node  $x_i$ . The key idea is to break up the minimax problem (3.10) into  $(N + 1)$  subproblems (3.11), each at every integration node  $x_i$ . The sought optimality measure for each integration node  $x_i$  can be stated as follows:

$$\text{Find } \alpha_i^* = \operatorname{argmin}_{\alpha > -1/2} \left| E_N^{(\alpha)}(x_i, \xi_i) \right|, \quad -1 < \xi_i < 1; \quad 0 \leq i \leq N. \quad (3.11)$$

Here  $\alpha_i^*$  is the optimal Gegenbauer parameter which minimizes the magnitude of the quadrature error at the integration node  $x_i$ , regardless of the magnitude of the  $(N + 1)$ th derivative of the integrand function  $f$ . Therefore, the construction of the numerical quadrature takes on a pointwise approach, where the corresponding unknown variables  $\xi_i = \xi_i(x_i)$  of the integration nodes  $x_i$  are treated as scalar numbers  $-1 < \xi_i < 1$ . In contrast, the previous approach seeks a sole optimal value  $\alpha^*$ , which minimizes the maximum of the magnitude of the quadrature error over  $\xi \in (-1, 1)$ . This in turn implies that the unknown variable  $\xi = \xi(x)$  is allowed to vary over the entire integration domain, and the function  $|f^{(N+1)}(\xi(x))|^p$  must be integrated throughout the whole interval  $[-1, 1]$ .

## Chapter 3

---

Let  $E_N^{(\alpha)}(x_i, \xi_i) = \psi(\xi_i)\eta_{i,N}(\alpha)$ , where  $\psi(\xi_i) = f^{(N+1)}(\xi_i)/(N+1)!$ ,  $\eta_{i,N}(\alpha) = \int_{-1}^{x_i} C_{N+1}^{(\alpha)}(x)dx/K_{N+1}^{(\alpha)}$ ;  $i = 0, \dots, N$ . Then

$$\left| E_N^{(\alpha)}(x_i, \xi_i) \right| = |\psi(\xi_i)| |\eta_{i,N}(\alpha)| \quad \forall 0 \leq i \leq N.$$

Using the identity  $\min cf(x) = c \min f(x) \forall c > 0$ , Problem (3.11) can be reduced to the following simple problem:

$$\text{Find } \alpha_i^* = \underset{\alpha > -1/2}{\operatorname{argmin}} |\eta_{i,N}(\alpha)| \quad \forall 0 \leq i \leq N. \quad (3.12)$$

It can be easily shown that  $\eta_{i,N}(\alpha)$  is a smooth function; therefore, one can exploit a second-order line search method to solve the problem, and obtain a rapid convergence to the optimal  $\alpha_i^*$  values. Since  $|\eta_{i,N}(\alpha)|$  is a nonsmooth function for each  $0 \leq i \leq N$ , one can readily recast Problem (3.12) as the following equivalent constrained minimization problem:

$$\text{Find } \alpha_i^* = \underset{\alpha > -1/2}{\operatorname{argmin}} \eta_{i,N}^2(\alpha), \quad 0 \leq i \leq N, \quad (3.13)$$

where the cost function is a smooth function. Problem (3.13) can be further converted into an unconstrained one-dimensional minimization problem using the change of variable (3.9). We notice here that the numerical scheme automatically constructs the optimal Gegenbauer quadrature using information only from the set of integration nodes  $\{x_i\}_{i=0}^N$ . In contrast, the construction of the optimal Gegenbauer quadrature presented in (El-Hawary et al., 2000) is function problematic, i.e., the quadrature construction method alters with the change of the underlying integrand function. We shall refer to the optimal GIM and the optimal Gegenbauer quadrature established through the solution of Problem (3.13) as the P-matrix and the P-matrix quadrature (PMQ), respectively.

In the standard method for constructing the  $\hat{P}$ -matrix, the integrand  $f(x)$  is interpolated at the same set of GG integration nodes  $S_N^{(\alpha)}$ , and the definite integrations  $\mathcal{J} = (\int_{-1}^{x_0} f(x), \dots, \int_{-1}^{x_N} f(x))^T$  are typically carried out by multiplying the constant  $\hat{P}$ -matrix with the column vector  $F$  of the integrand values at the GG points  $x_i \in S_N^{(\alpha)}$  as given by Equation (3.2). On the other hand, the design of the P-matrix is established by taking into account the effect of each integration node  $x_i$  separately on the truncation error, and minimizes the error optimally in the sense of solving Problem (3.13). As a result, this novel pointwise approach takes a different path for evaluating the required definite integrations. In particular, for each integration node  $x_i$ , an optimal Gegenbauer parameter  $\alpha_i^*$  is determined, and instead of interpolating the integrand  $f(x)$  at the set of interpolation points  $S_N^{(\alpha)}$ , the PMQ seeks a new GG points set as the optimal GG interpolation points

## Chapter 3

---

set corresponding to the integration node  $x_i$ . We shall denote these optimal GG interpolation nodes by  $z_{i,k}, k = 0, \dots, M$ , for some  $M \in \mathbb{Z}^+$ , and we shall call them the generalized/adjoint GG points, since they generally differ from the particular choice of the GG integration nodes  $\{x_i\}_{i=0}^N$ , and adjoin each integration node  $x_i$  in the construction of the PMQ.

The free choice of the number  $M$  of adjoint GG nodes  $z_{i,j}$  renders the P-matrix a rectangular matrix of size  $(N + 1) \times (M + 1)$  rather than a square matrix of size  $(N + 1)$ , as is typically the case with a conventional spectral integration matrix. Hence the optimal first-order P-matrix can be conveniently written as  $P^{(1)} = (p_{ik}^{(1)}(\alpha_i^*)), i = 0, \dots, N; k = 0, \dots, M$ , where  $p_{ik}^{(1)}(\alpha_i^*)$  are the matrix elements of the  $i$ th row obtained using the optimal value of  $\alpha_i^*$ . The definite integral  $\int_{-1}^{x_i} f(x)dx$  is then approximated by the optimal Gegenbauer quadrature as follows:

$$\int_{-1}^{x_i} f(x)dx \approx \sum_{k=0}^M p_{ik}^{(1)}(\alpha_i^*)f(z_{ik}) \quad \forall i = 0, \dots, N. \quad (3.14)$$

It can be shown that the Gegenbauer approximation of the definite integrals  $J$  using the PMQ can be described in matrix form through the Hadamard product (entrywise product) instead of the usual matrix-vector multiplication, as we shall discuss later in Section 3.2.8. In the following section, we shall describe the method of constituting the elements of the P-matrix, and analyze the PMQ truncation error.

### 3.2.2 Generation of the P-matrix and Error Analysis

The following theorem describes the construction of the P-matrix elements, and highlights the truncation error of the resulting PMQ:

**Theorem 3.2.3.** *Let*

$$S_{N,M} = \{z_{i,k} | C_{M+1}^{(\alpha_i^*)}(z_{i,k}) = 0, i = 0, \dots, N; k = 0, \dots, M\}, \quad (3.15)$$

*be the generalized/adjoint set of GG points, where  $\alpha_i^*$  are the optimal Gegenbauer parameters in the sense that*

$$\alpha_i^* = \underset{\alpha > -1/2}{\operatorname{argmin}} \eta_{i,M}^2(\alpha), \quad (3.16)$$

$$\eta_{i,M}(\alpha_i^*) = \int_{-1}^{x_i} C_{M+1}^{(\alpha_i^*)}(x)dx / K_{M+1}^{(\alpha_i^*)}; \quad (3.17)$$

$$K_{M+1}^{(\alpha_i^*)} = 2^M \frac{\Gamma(M + \alpha_i^* + 1)\Gamma(2\alpha_i^* + 1)}{\Gamma(M + 2\alpha_i^* + 1)\Gamma(\alpha_i^* + 1)}. \quad (3.18)$$

## Chapter 3

---

Moreover, let  $f(x) \in C^\infty[-1, 1]$  be approximated by the Gegenbauer polynomials expansion series such that the Gegenbauer coefficients are computed by interpolating the function  $f(x)$  at the adjoint GG points  $z_{i,k} \in S_{N,M}$ . Then there exist a matrix  $P^{(1)} = (p_{ij}^{(1)})$ ,  $i = 0, \dots, N$ ;  $j = 0, \dots, M$ ; and some numbers  $\xi_i \in [-1, 1]$  satisfying

$$\int_{-1}^{x_i} f(x) dx = \sum_{k=0}^M p_{ik}^{(1)}(\alpha_i^*) f(z_{i,k}) + E_M^{(\alpha_i^*)}(x_i, \xi_i), \quad (3.19)$$

where

$$p_{ik}^{(1)}(\alpha_i^*) = \sum_{j=0}^M (\lambda_j^{(\alpha_i^*)})^{-1} \omega_k^{(\alpha_i^*)} C_j^{(\alpha_i^*)}(z_{i,k}) \int_{-1}^{x_i} C_j^{(\alpha_i^*)}(x) dx, \quad (3.20)$$

$$(\omega_k^{(\alpha_i^*)})^{-1} = \sum_{j=0}^M (\lambda_j^{(\alpha_i^*)})^{-1} (C_j^{(\alpha_i^*)}(z_{i,k}))^2, \quad (3.21)$$

$$\lambda_j^{(\alpha_i^*)} = \frac{2^{2\alpha_i^* - 1} j! \Gamma^2(\alpha_i^* + \frac{1}{2})}{(j + \alpha_i^*) \Gamma(j + 2\alpha_i^*)}, \quad (3.22)$$

$$E_M^{(\alpha_i^*)}(x_i, \xi_i) = \frac{f^{(M+1)}(\xi_i)}{(M+1)!} \eta_{i,M}(\alpha_i^*). \quad (3.23)$$

*Proof.* See Appendix 3.B. □

Theorem 3.2.3 shows that the PMQ adapts to deal with any arbitrary sets of integration nodes. In particular, for a certain integration node  $x_i$ , the novel approach of the PMQ method determines an optimal Gegenbauer parameter  $\alpha_i^*$ , which optimally breaks down the quadrature error in the sense of solving Problem (3.13). The construction of the PMQ is then carried out through interpolation at the optimal set of adjoint GG points  $z_{i,k} \in S_{N,M}$  corresponding to the integration node  $x_i$ . Hence the approximation of the definite integrations  $\int_{-1}^{x_i} f(x) dx$  of a smooth function  $f(x)$  is carried out through expansions in distinct Gegenbauer polynomials associated with finitely many optimal Gegenbauer parameters  $\alpha_i^*$  corresponding to the integration nodes  $x_i$ . In contrast, typical Gegenbauer quadrature methods employ a unique Gegenbauer expansion series with a fixed  $\alpha$  parameter. This feature distinguishes the PMQ from other spectral quadrature methods in the literature. The necessary steps for constructing the PMQ are conveniently described in Figure 3.1. One of the main contributions of the novel PMQ lies in the achievement of approximations of higher-order than those obtained by the standard Chebyshev, Legendre, and Gegenbauer expansion methods at least for a small range of the number of spectral expansion terms, as we shall demonstrate later via an extensive set of test problems worked through in Section 3.3. Moreover, this numerical technique establishes high-precision approximations

## Chapter 3

---

for any arbitrary sets of integration nodes  $x_i$ , where the Gegenbauer quadrature scheme no longer depends on the specific type of the integration nodes.

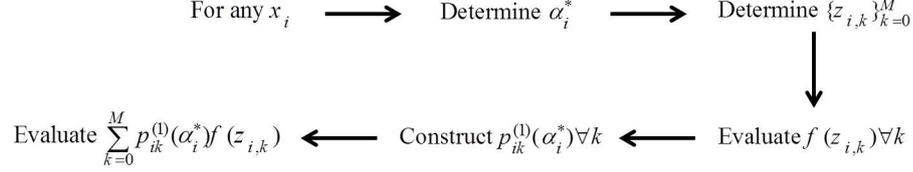


Figure 3.1: The steps for evaluating the definite integral  $\int_{-1}^{x_i} f(x)dx$  of a given function  $f(x) \in C^\infty[-1, 1]$  by using Theorem 3.2.3. The figure shows that instead of strictly using the set of integration nodes  $\{x_i\}_{i=0}^N = S_N^{(\alpha)}$  that is the same as the set of interpolation points required for constructing the Gegenbauer quadrature, one chooses any arbitrary set of integration nodes  $\{x_i\}_{i=0}^N$ . For a particular integration node  $x_i$ , the PMQ determines the optimal Gegenbauer parameter  $\alpha_i^*$  in the sense of minimizing the square of the  $\eta$ -function,  $\eta_{i,M}^2(\alpha)$ . The PMQ then employs the adjoint GG nodes  $z_{i,k}$  corresponding to the integration node  $x_i$  as the optimal set of interpolation points, and evaluates the integrand  $f(x)$  at these optimal points. The Gegenbauer quadrature method proceeds by constructing the  $i$ th row of the P-matrix,  $(p_{i,0}^{(1)}(\alpha_i^*), p_{i,1}^{(1)}(\alpha_i^*), \dots, p_{i,M}^{(1)}(\alpha_i^*))$ , and evaluates the definite integral  $\int_{-1}^{x_i} f(x)dx$  as stated by Formula (3.14).

### 3.2.3 Polynomial Integration

Since integration of polynomials appears frequently in many scientific areas, it is important to analyze the PMQ error for general polynomials. The following corollary highlights the PMQ truncation error for polynomials of any arbitrary degree  $n$ :

**Corollary 3.2.4** (Polynomial integration). *The PMQ (3.14) is exact for all polynomials  $h_n(x)$  of arbitrary degree  $n$  for any set of integration nodes  $\{x_i\}_{i=0}^N \subset [-1, 1]$ ;  $M \geq n$ .*

*Proof.* The truncation error of the PMQ as stated by Equation (3.23) is given by

$$E_M^{(\alpha_i^*)}(x_i, \xi_i) = \frac{h_n^{(M+1)}(\xi_i)}{(M+1)!} \eta_{i,M}(\alpha_i^*), \quad (3.24)$$

which is identically zero for all  $M \geq n$ .  $\square$

Corollary 3.2.4 shows that the round-off error is the only source of error arising from the calculation of the PMQ for polynomials of any arbitrary degree  $n$  if  $M \geq n$ . In contrast, the optimal QMQ is not exact for polynomials of degree

## Chapter 3

---

$n$  unless the number of integration/interpolation nodes  $N \geq n$ . In the following section we shall investigate the convergence rate of the PMQ for smooth functions, and the bounds on its truncation error.

### 3.2.4 Error Bounds and Convergence Rate

To determine an error bound for the truncation error of the PMQ for the asymptote  $M \rightarrow \infty$ , and to analyze the convergence rate of the quadrature, we require the following two lemmas:

**Lemma 3.2.5.** *The maximum value of the Gegenbauer polynomials  $C_M^{(\alpha)}(x)$  generated by Equations (3.A.4) is less than or equal to 1 for all  $\alpha \geq 0; M \geq 0$ , and of order  $M^{-\alpha}$  for all  $-1/2 < \alpha < 0; M \gg 1$ .*

*Proof.* See (Elgindy and Smith-Miles, 2013c). □

**Lemma 3.2.6.** *For a fixed  $\alpha > -1/2$ , the factor  $1/((M+1)!K_{M+1}^{(\alpha)})$  is of order  $1/(M^{\frac{1}{2}-\alpha}(2M/e)^M)$  for large values of  $M$ , where  $e$  is the base of the natural logarithm.*

*Proof.* See (Elgindy and Smith-Miles, 2013c). □

Lemma 3.2.5 highlights the magnitude of the Gegenbauer polynomials of increasing orders, while Lemma 3.2.6 is significant for analyzing the convergence rate of the PMQ. In fact, Lemma 3.2.6 illustrates that for a fixed  $\alpha > -1/2$ , the factor  $1/((M+1)!K_{M+1}^{(\alpha)})$  decays exponentially faster than any finite power of  $1/M$ . This error factor is the major element in damping the PMQ truncation error as the value of  $M$  increases. The following theorem estimates the bound on the truncation error for increasing values of  $M$ :

**Theorem 3.2.7** (Error bounds). *Assume that  $f(x) \in C^\infty[-1, 1]$ , and  $\max_{|x| \leq 1} |f^{(M+1)}(x)| \leq A \in \mathbb{R}^+$ , for some number  $M \in \mathbb{Z}^+$ . Moreover, let  $\int_{-1}^{x_i} f(x)dx$  be approximated by the PMQ (3.14) up to the  $(M+1)^{\text{th}}$  Gegenbauer expansion term, for each integration node  $x_i, i = 0, \dots, N$ . Then there exist some positive constants  $D_1^{(\alpha_i^*)}; D_2^{(\alpha_i^*)}$  independent of  $M$  such that the magnitude of the PMQ truncation error  $E_M^{(\alpha_i^*)}$  is bounded by the following inequality:*

$$\left| E_M^{(\alpha_i^*)} \right| \leq \begin{cases} B_1^{(\alpha_i^*)} \left(\frac{e}{2}\right)^M \frac{1+x_i}{M^{M+1/2-\alpha_i^*}}, & \alpha_i^* \geq 0, \\ B_2^{(\alpha_i^*)} \left(\frac{e}{2}\right)^M \frac{1+x_i}{M^{M+1/2}}, & \alpha_i^* < 0, \end{cases} \quad (3.25)$$

for all  $0 \leq i \leq N$ , as  $M \rightarrow \infty$ , where  $B_1^{(\alpha_i^*)} = AD_1^{(\alpha_i^*)}; B_2^{(\alpha_i^*)} = B_1^{(\alpha_i^*)} D_2^{(\alpha_i^*)}$ .

## Chapter 3

---

*Proof.* See Appendix 3.C. □

Theorem 3.2.7 shows that the error bound decays exponentially faster than any finite power of  $1/M$ , and the PMQ exhibits rapid spectral convergence for increasing values of  $M$  regardless of the number of integration nodes  $N$ . In fact, a similar analysis carried on the standard QMQ shows that for a certain integration node  $x_i \in S_N^{(\alpha)}$  with  $\alpha \geq 0$ , the bound on the PMQ truncation error decays faster than that of the QMQ by a factor  $\mathcal{R}_1$ :

$$\mathcal{R}_1 = \begin{cases} \hat{B}_1^{(\alpha, \alpha_i^*)} \left(\frac{e}{2}\right)^{M-N} \frac{N^{N+\frac{1}{2}-\alpha}}{M^{M+\frac{1}{2}-\alpha_i^*}}, & \alpha_i^* \geq 0, \\ \hat{B}_2^{(\alpha, \alpha_i^*)} \left(\frac{e}{2}\right)^{M-N} \frac{N^{N+\frac{1}{2}-\alpha}}{M^{M+\frac{1}{2}}}, & \alpha_i^* < 0, \end{cases} \quad (3.26)$$

for some constants  $\hat{B}_1^{(\alpha, \alpha_i^*)}; \hat{B}_2^{(\alpha, \alpha_i^*)}$ , independent of the numbers  $N; M$ , assuming  $M > N$ . One can show further that for a certain integration node  $x_i \in S_N^{(\alpha^*)}$  with  $\alpha^* \geq 0$ , and under the same assumption  $M > N$ , there exist some constants  $\bar{B}_1^{(\alpha^*, \alpha_i^*)}; \bar{B}_2^{(\alpha^*, \alpha_i^*)}$ , independent of the numbers  $N; M$ , such that the bound on the PMQ truncation error decays faster than that of the optimal QMQ by a factor  $\mathcal{R}_2$ :

$$\mathcal{R}_2 = \begin{cases} \bar{B}_1^{(\alpha^*, \alpha_i^*)} \left(\frac{e}{2}\right)^{M-N} \frac{N^{N+\frac{1}{2}-\alpha^*}}{M^{M+\frac{1}{2}-\alpha_i^*}}, & \alpha_i^* \geq 0, \\ \bar{B}_2^{(\alpha^*, \alpha_i^*)} \left(\frac{e}{2}\right)^{M-N} \frac{N^{N+\frac{1}{2}-\alpha^*}}{M^{M+\frac{1}{2}}}, & \alpha_i^* < 0. \end{cases} \quad (3.27)$$

Hence the rectangular form of the P-matrix allows for faster convergence rates by increasing the number of columns ( $M + 1$ ) without increasing the number of integration nodes ( $N + 1$ ); cf. Table 3.4 in Section 3.3. In contrast, the accuracy of the QMQ cannot be improved unless we increase the number of integration/interpolation points ( $N + 1$ ), since the Q-matrix is a square matrix of size ( $N + 1$ ).

### 3.2.5 Convergence of the PMQ to the Chebyshev Quadrature in the $L^\infty$ -norm

A well-known result in approximation theory states that for sufficiently large spectral expansion terms, the truncated expansions in Chebyshev polynomials are optimal for the  $L^\infty$ -norm approximations of smooth functions (Balachandran et al., 2009), while the truncated expansions in Legendre polynomials are

## Chapter 3

---

optimal for the  $L^2$ -norm approximations of smooth functions (Fornberg, 1996). Therefore, it is natural to acknowledge that the Chebyshev quadrature is optimal for the  $L^\infty$ -norm approximations of the definite integrals of smooth functions, while the Legendre quadrature is optimal for the  $L^2$ -norm approximations of the definite integrals of smooth functions. The following theorem shows that the PMQ constructed by solving Problem (3.13) converges to the optimal Chebyshev quadrature in the  $L^\infty$ -norm, for a large-scale number of expansion terms.

**Theorem 3.2.8** (Convergence of the PMQ). *Assume that  $f(x) \in C^\infty[-1, 1]$ , and  $\max_{|x| \leq 1} |f^{(M+1)}(x)| \leq A \in \mathbb{R}^+$ , for some number  $M \in \mathbb{Z}^+$ . Moreover, let  $\int_{-1}^{x_i} f(x) dx$  be approximated by the PMQ (3.14) up to the  $(M+1)^{\text{th}}$  Gegenbauer expansion term, for each integration node  $x_i, i = 0, \dots, N$ . Then the PMQ converges to the optimal Chebyshev quadrature in the  $L^\infty$ -norm as  $M \rightarrow \infty$ .*

*Proof.* See Appendix 3.D. □

Theorem 3.2.8 shows that the PMQ is the ideal Gegenbauer quadrature from three perspectives: (i) The PMQ takes full advantage of the parent family of the Gegenbauer polynomials in the approximation of the definite integrals of smooth functions for the small range of the spectral expansion terms, where the precision of the optimal Gegenbauer quadrature can exceed those obtained by the standard Chebyshev, Legendre, and Gegenbauer quadratures; cf. (Elgindy and Smith-Miles, 2013c) and Section 3.3 in this chapter. In fact, in this case, it can be demonstrated that the PMQ constructed using negative and nonnegative optimal Gegenbauer parameters  $\alpha_i^*$  may produce parallel higher-order approximations, and the Gegenbauer polynomials are generally more effective than the standard Chebyshev and Legendre polynomials; cf. Section 3.3. (ii) The PMQ takes full advantage of the optimal Chebyshev polynomials in the  $L^\infty$ -norm approximation of definite integrals of smooth functions for a sufficiently large range of the spectral expansion terms. (iii) The previous two attractive features of the PMQ can be accomplished regardless of the number of integration nodes  $N$ .

### 3.2.6 Determining the Interval of Uncertainty for the Optimal Gegenbauer Parameters of the P-matrix for Small/Medium Range Expansions

In the previous section, we proved that the PMQ converges to the optimal Chebyshev quadrature in the  $L^\infty$ -norm for large-scale number of Gegenbauer expansion terms. In this section, we attempt to determine the interval of uncertainty where the optimal Gegenbauer parameters  $\alpha_i^*$  can be found. This matter is crucial for the line search method, since a tight search interval embedding the optimal values

## Chapter 3

---

of  $\alpha_i^*$  can reduce the required computations, and the calculation time. A straightforward analysis on the behavior of the Gegenbauer polynomials shows that the values of  $\alpha_i^* \approx -0.5$  may break the numerical stability of the PMQ scheme, since the Gegenbauer polynomials of increasing orders grow rapidly as  $\alpha_i^* \rightarrow -0.5$ , as is evident from Equation (3.A.3), and only suitable negative values of  $\alpha_i^*$  are to be employed to produce better approximations. Hence the left endpoint of the search interval must not be too close to the value  $-0.5$ . Now we investigate the potential search interval along the positive values of  $\alpha$ . The following lemma is crucial in determining the candidate interval of uncertainty:

**Lemma 3.2.9.** *For a fixed number  $j \in \mathbb{Z}^+ \cup \{0\}$ , the normalization factor  $\lambda_j^{(\alpha)} \rightarrow 0$  as  $\alpha \rightarrow \infty$ .*

*Proof.* See Appendix 3.E. □

Hence the magnitude of the normalization factor  $\lambda_j^{(\alpha)}$ , for each  $j$ , diminishes for increasing values of  $\alpha$ . In fact, this behavior is foreseen in view of the nature of the Gegenbauer weight function  $w^{(\alpha)}(x)$ , which narrows for increasing values of  $\alpha$ ; cf. Figure 3.2. Since the Gegenbauer polynomials are level functions for  $\alpha \geq 0$ , as is evident from Lemma 3.2.5, i.e. they oscillate smoothly between  $+1$  and  $-1$  in the interval  $[-1, 1]$ , then the maximum value of the definite integral (3.A.6) is utterly dominated by the value of the weight function, which collapses for  $\alpha \gg 1$ . This behavior of the Gegenbauer weight function gives rise to the following important theorem:

## Chapter 3

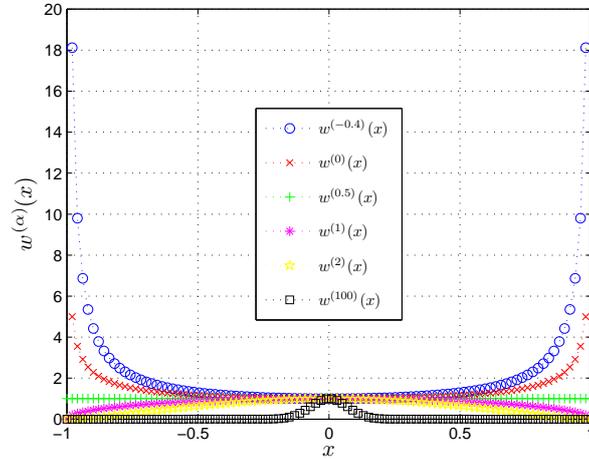


Figure 3.2: The profile of the Gegenbauer weight function  $w^{(\alpha)}(x)$  for  $\alpha = -0.4, 0, 0.5, 1, 2; 100$ . Clearly the weight function dies out near the boundaries  $x = \pm 1$ , and the function becomes nonzero only on a small subdomain centered at the middle of the interval  $x = 0$  for increasing values of  $\alpha$ .

**Theorem 3.2.10.** *For a fixed number  $(M + 1)$  of Gegenbauer expansion terms, the elements of the  $i$ th row of the  $P$ -matrix vanish very rapidly as  $\alpha_i^* \rightarrow \infty$ , for each  $i$ .*

*Proof.* See Appendix 3.F. □

Theorem 3.2.10 shows the significant effect of applying the PMQ using large and positive values of  $\alpha_i^*$ . In particular, we notice two major results: (i) Firstly, although the elements of the  $P$ -matrix converge to zero for increasing values of  $\alpha_i^*$  theoretically for a fixed  $M$ , in practice, the operation of multiplication of the very large number  $(\lambda_s^{(\alpha_i^*)})^{-1}$  with the very small number  $\mathcal{K}_{i,s}$  for increasing values of  $\alpha_i^*$  is prone to large round-off error, causing degradation in the observed precision. (ii) Secondly, and most importantly, the PMQ may become sensitive to round-off errors for positive and large values of the parameter  $\alpha$  due to the narrowing effect of the Gegenbauer weight function  $w^{(\alpha)}(x)$ , forcing the PMQ to become more extrapolatory, i.e., the PMQ may rely heavily on the behavior of the integrand function central to the sampled interval for anticipating its behavior closer to the boundaries  $x = \pm 1$ . The narrowing behavior of the weight function commences closely from the values of  $\alpha = 1; 2$ , as shown by Figure 3.2. This analysis suggests choosing the uncertainty interval  $(-0.5 + \varepsilon, r)$ , to maintain higher-order approximations, and to obtain satisfactory results, where  $\varepsilon$  is a relatively small and positive parameter;  $r \in [1, 2]$ . The left limit acts as a barrier preventing

## Chapter 3

---

the numerical instability caused by the growth of the Gegenbauer polynomials of increasing orders for  $\alpha \approx -0.5$ , while the right limit acts as another barricade hampering the extrapolatory effect of the PMQ caused by the narrowing behavior of the Gegenbauer weight function for increasing values of  $\alpha$ .

### 3.2.7 Substantial Advantages of the Gegenbauer Collocation Methods Endowed with the PMQ

It can be shown that the integration nodes  $\{x_i\}_{i=0}^N$  are typically the same as the collocation points in a Gegenbauer collocation integration method discretizing various continuous mathematical models. To illustrate this point, consider for simplicity, and without loss of generality, the following linear TPBVP:

$$y''(x) = f(x)y'(x) + g(x)y(x) + r(x), \quad 0 \leq x \leq 1, \quad (3.28a)$$

with the Dirichlet boundary conditions

$$y(0) = \beta; \quad y(1) = \gamma. \quad (3.28b)$$

Direct integration converts the problem into the following integral counterpart:

$$y(x) = \int_0^x \int_0^{t_2} ((g(t_1) - F(t_1))y(t_1) + r(t_1))dt_1dt_2 + \int_0^x f(t)y(t)dt + (c_1 - \beta f_0)x + c_2, \quad (3.29)$$

where  $F \equiv f'$ ;  $f_0 = f(0)$ . The constants  $c_1$  and  $c_2$  are the integration constants determined through the boundary conditions (3.28b). Hence the TPBVP can be written at the collocation points  $\{x_i\}_{i=0}^N$  as

$$y(x_i) - \int_0^{x_i} \int_0^{t_2} ((g(t_1) - F(t_1))y(t_1) + r(t_1))dt_1dt_2 - \int_0^{x_i} f(t)y(t)dt + (\beta f_0 - c_1)x_i - c_2 = 0, \quad (3.30)$$

where the set of integration nodes  $\{x_i\}_{i=0}^N$  is the same as the collocation points set. To convert the integral equation into an algebraic equation, one needs some expressions for approximating the integral operators involved in the integral equation. In a spectral collocation method, this operation is conveniently carried out through the spectral integration matrices. While traditional spectral methods demand that the number of spectral expansion terms ( $N + 1$ ) required for the construction of the spectral differentiation/integration matrix be exactly the same as the number of collocation points; cf. (El-Gendi, 1969; Elbarbary, 2007; Fornberg, 1990; Ghoreishi and Hosseini, 2004; Gong et al., 2009; Paraskevopoulos, 1983; Ross and Fahroo, 2002; Weideman and Reddy, 2000), Theorem 3.2.3 shows

## Chapter 3

---

that the choice of the number of Gegenbauer expansion terms  $(M + 1)$  is completely free. Consequently, two substantial advantages can be drawn from the application of the PMQ in a Gegenbauer collocation integration method:

- (i) For any small number of collocation points  $(N + 1)$ , the Gegenbauer collocation method can boost the precision of the approximate solutions by increasing the number of the PMQ expansion terms  $(M + 1)$  without increasing the value of  $N$ . Consequently, one can obtain higher-order approximations of the solutions of complex mathematical problems without increasing the number of collocation points. The reader may consult the recent paper of Elgindy and Smith-Miles (2013c) for clear examples on this important result.
- (ii) For any large number of collocation points  $(N + 1)$ , the Gegenbauer collocation method can produce very precise approximations to the smooth solutions of the mathematical model in a very short time by restricting the value of  $M$  to accepting only small values—typically a value of  $M$  in the range  $14 \leq M \leq 16$  is usually satisfactory and sufficient for producing higher-order approximations for many problems and applications; cf. (Elgindy and Smith-Miles, 2013c; Elgindy et al., 2012).

### 3.2.8 The Matrix Form Gegenbauer Approximation of Definite Integrations

To describe the approximations of the definite integrals  $\int_{-1}^{x_i} f(x)dx, i = 0, \dots, N$ , of a function  $f(x) \in C^\infty[-1, 1]$  in matrix form using the P-matrix, let  $P^{(1)}$  be the first-order rectangular P-matrix of size  $(N + 1) \times (M + 1)$ , where  $M$  denotes the highest degree of the Gegenbauer polynomial employed in the PMQ, and set  $P^{(1)}$  in the block matrix form  $P^{(1)} = (P_0^{(1)} P_1^{(1)} \dots P_N^{(1)})^T, P_i^{(1)} = (p_{i,0}^{(1)}, p_{i,1}^{(1)}, \dots, p_{i,M}^{(1)}); i = 0, \dots, N$ . Also let  $V$  be a rectangular matrix of size  $(M + 1) \times (N + 1)$  such that  $V = (V_0 V_1 \dots V_N), V_i = (f(z_{i,0}), f(z_{i,1}), \dots, f(z_{i,M}))^T, i = 0, \dots, N; f(z_{ij})$  is the function  $f$  calculated at the adjoint GG points  $z_{i,j} \in S_{N,M}$ . Then the approximations of the required definite integrations of  $f(x)$  using the P-matrix are given by

$$\left( \int_{-1}^{x_0} f(x)dx, \int_{-1}^{x_1} f(x)dx, \dots, \int_{-1}^{x_N} f(x)dx \right)^T \approx P^{(1)} \circ V^T, \quad (3.31)$$

where  $\circ$  is the Hadamard product with the elements of  $P^{(1)} \circ V^T$  given by

$$(P^{(1)} \circ V^T)_i = P_i^{(1)} \cdot V_i = \sum_{j=0}^M p_{i,j}^{(1)} f(z_{i,j}), \quad i = 0, \dots, N. \quad (3.32)$$

## Chapter 3

---

To calculate the  $n$ -fold definite integrals  $I_i^{(n)}$  of the function  $f$  defined by

$$I_i^{(n)} = \int_{-1}^{x_i} \int_{-1}^{t_{n-1}} \cdots \int_{-1}^{t_2} \int_{-1}^{t_1} f(t_0) dt_0 dt_1 \cdots dt_{n-2} dt_{n-1} \quad \forall 0 \leq i \leq N,$$

we can use Cauchy's formula which reduces certain iterated integrals to a single integral as follows:

$$\int_{-1}^{x_i} \int_{-1}^{t_{n-1}} \cdots \int_{-1}^{t_2} \int_{-1}^{t_1} f(t_0) dt_0 dt_1 \cdots dt_{n-2} dt_{n-1} = \frac{1}{(n-1)!} \int_{-1}^{x_i} (x_i - t)^{n-1} f(t) dt.$$

Hence

$$(I_0^{(n)}, I_1^{(n)}, \dots, I_N^{(n)})^T \approx P^{(n)} \circ V^T, \quad (3.33)$$

where  $P^{(n)} = (p_{i,j}^{(n)})$  is the  $n^{\text{th}}$ -order P-matrix with the elements

$$p_{i,j}^{(n)} = \frac{(x_i - z_{i,j})^{n-1}}{(n-1)!} p_{i,j}^{(1)}, \quad i = 0, \dots, N; j = 0, \dots, M \forall x \in [-1, 1]. \quad (3.34)$$

For the integration over the interval  $[0, 1]$ , Equation (3.34) is replaced with

$$p_{i,j}^{(n)} = \frac{(x_i - z_{i,j})^{n-1}}{2^n (n-1)!} p_{i,j}^{(1)}, \quad i = 0, \dots, N; j = 0, \dots, M. \quad (3.35)$$

Hence the P-matrices of higher-orders can be quickly generated from the first-order P-matrix. In the next section, we present some efficient computational algorithms for the construction of the P-matrix.

### 3.2.9 Computational Algorithms

In this section, we turn our attention to the development of two efficient algorithms for constructing the P-matrix. The proposed algorithms take into account the following important elements: (i) The approximations of the definite integrals of smooth functions are required to be in the  $L^2$ -norm. (ii) For the small range number of the number of Gegenbauer expansion terms ( $M + 1$ ), the optimal Gegenbauer quadrature can produce higher-order approximations than the standard Chebyshev, Legendre, and Gegenbauer polynomials quadratures, as we shall demonstrate later in Section 3.3. (iii) The PMQ may become sensitive to round-off errors for positive and large values of the parameter  $\alpha$  due to the narrowing nature of the Gegenbauer weight function  $w^{(\alpha)}(x)$ . (iv) The values of  $\alpha_i^* \approx -0.5$  may break the numerical stability of the PMQ scheme, since the Gegenbauer polynomials of increasing orders grow rapidly as  $\alpha_i^* \rightarrow -0.5$ . (v) For large values of  $M$ , the truncated expansions in Legendre polynomials are optimal

## Chapter 3

---

for the  $L^2$ -norm approximations of smooth functions. That is, for a large value of  $M$ , there exists an integer number  $M_{\max} \in \mathbb{Z}^+$  such that the Legendre polynomial  $L_M(x)$  is the best possible polynomial approximant in the  $L^2$ -norm for all  $M \geq M_{\max}$ . (vi) For large values of  $M$ , the PMQ automatically converges to the Chebyshev quadrature. Therefore, we need a convenient method for forcing the computational algorithm to construct the rectangular/square Legendre matrix, and obtain the Legendre quadrature instead. (vii) For large values of  $M$ , the Legendre polynomials  $L_M(x)$  are less sensitive to the small perturbations in the argument  $x$  than the Gegenbauer polynomials  $C_M^{(\alpha)}(x)$  associated with negative Gegenbauer parameters  $\alpha$ . To clarify this last item, consider the Gegenbauer polynomial  $C_{150}^{(-1/4)}(x)$  of degree 150, and associated with the Gegenbauer parameter  $\alpha = -1/4$ . Evaluating this Gegenbauer polynomial at  $x = 1/2$  in exact arithmetic using MATHEMATICA 8 software Version 8.0.4.0 yields the exact value 5.754509478448837 accurate to 16 decimal digits. In practice, and working in a floating-point arithmetic environment, one should expect a perturbation in the value of the argument  $x$ . Evaluating the same polynomial at  $x = 0.5001$  with a small perturbation of  $10^{-4}$  in the value of its argument  $x$  gives the exact value 5.766713747271598 accurate to 16 decimal digits, with an absolute error of approximately 0.0122. Hence a slight change in the argument of the Gegenbauer polynomials of higher orders, and associated with negative  $\alpha$  values, may ruin the spectral convergence properties of the Gegenbauer quadrature method. It can be noticed however that this sensitivity is not of great concern for the Chebyshev and Legendre polynomials. For instance, for the same practical example, the absolute errors associated with the Chebyshev and Legendre polynomials of orders 150 are approximately  $1.5001 \times 10^{-4}$ ;  $3.2286 \times 10^{-4}$ , respectively. This shows that the Chebyshev and Legendre polynomials are very attractive for large expansions of the spectral expansion terms. In the argument above, we highlighted only the ill-conditioning of the Gegenbauer polynomials of higher orders and associated with negative Gegenbauer parameters, since the Gegenbauer polynomials of higher orders and associated with positive Gegenbauer parameters are generally well-conditioned, and their well-conditioning increases for increasing values of  $\alpha$ . For instance, for the same example above and using  $\alpha = 1/4$ ;  $3/4$  instead of  $-1/4$ , we obtain the absolute errors of  $5.8508 \times 10^{-4}$ ;  $1.5056 \times 10^{-4}$ , respectively.

These influential factors suggest that the PMQ constructed naively by solving Problem (3.13) without taking into consideration the aforementioned elements is practically not optimal. These important elements have motivated us to develop two efficient and robust algorithms for the construction of the P-matrix for general symmetric/nonsymmetric sets of integration nodes, where the strengths of the Chebyshev, Legendre, and Gegenbauer polynomials are exploited. In the following section, we describe our first construction algorithm for general non-

## Chapter 3

---

symmetric integration points.

### 3.2.9.1 The Nonsymmetric Integration Points Case

To maintain consistency with the aforementioned factors, Algorithm 2.1 (see Appendix 3.G) implements the Gegenbauer polynomial expansion over a small/medium scale of  $M$ , and implements the Legendre polynomial expansion for increasing values of  $M$ . Here the user inputs three integer numbers  $N, M; M_{\max}$ . The former two define the size of the P-matrix,  $P^{(1)} \in \mathbb{R}^{(N+1) \times (M+1)}$ , while the last,  $M_{\max}$ , defines the maximum number of  $M$  at which the algorithm transits into the implementation of the Legendre polynomial expansion. The user also inputs the right endpoint  $r$  of the search interval for the optimal Gegenbauer parameters, the set of integration points  $\{x_i\}_{i=0}^N$ , and a relatively small positive number  $\varepsilon$ . Typically  $r \in [1, 2]$  is a suitable choice as discussed before in Section 3.2.6. In Step 1, the algorithm checks whether the number  $M$  is greater than the prescribed number  $M_{\max}$ . If the condition is satisfied and  $M = N$ , then the algorithm constructs the square Legendre matrix  $\hat{P}$  (with  $\alpha = 0.5$ ). Otherwise the algorithm constructs a modified Legendre matrix of rectangular form. If  $M \leq M_{\max}$ , the algorithm constructs the  $i$ th row of the P-matrix, for each  $0 \leq i \leq N$  in Steps 2-7. Here we would like to stress the importance of Step 5 of the algorithm. In practice, and as we have discussed before, it is seldom advantageous to choose the values of  $\alpha_i^* \in (-0.5, -0.5 + \varepsilon)$ , as the Gegenbauer polynomial  $C_M^{(\alpha_i^*)}(x)$  of fixed degree  $M$  grows rapidly as  $\alpha_i^* \rightarrow -0.5$ , and the round-off error dominates the calculations. Therefore Step 5 is necessary for improving the values of the critical  $\alpha$  values obtained from Step 4 in the cases where  $\alpha_i^* \in (-0.5, -0.5 + \varepsilon)$ . Here the algorithm technique is to choose the value of  $\alpha_i^*$  at the right limit of the critical interval  $(-0.5, -0.5 + \varepsilon)$ . The reason of this choice is to prevent any potential escalation in the PMQ truncation error for an arbitrary value of  $\alpha_i^*$  distant from this critical interval. The choice of  $\alpha_i^* = 0.5$  is another natural choice suitable as a viable alternative if the calculations are sensitive to the value of  $\varepsilon$ . To illustrate further, Step 5 of the algorithm works as a safeguard step in the cases where the solution of Problem (3.13) at a certain integration point  $x_i$  falls in the neighborhood of the boundary value  $\alpha = -0.5$ , and forces the algorithm to either choose  $\alpha_i^*$  at the right limit of this critical interval or apply Legendre polynomials to construct the corresponding  $i$ th row of the P-matrix. Finally, Step 8 of the algorithm outputs the constructed P-matrix and terminates the code. Notice here that the contribution of Step 1 of the algorithm is important, since the PMQ automatically converges to the Chebyshev quadrature for large values of  $M$  as proven by Theorem 3.2.8, which is not optimal in the sense of the  $L^2$ -norm. For this particular reason, Step 1 drives the algorithm to construct the optimal Legendre quadrature. Moreover, Step 1 provides the user with the convenience of applying

## Chapter 3

---

the Legendre polynomial expansions as soon as  $M$  exceeds the maximum value  $M_{\max}$ . Hence the convergence properties of the PMQ are the same as those of the Legendre polynomial quadratures for the values of  $M > M_{\max}$ .

### 3.2.9.2 The Symmetric Integration Points Case: A More Computationally Efficient Algorithm

For practical considerations, if the set of integration nodes  $\{x_i\}_{i=0}^N$  is symmetric, then the efficiency of Algorithm 2.1 can be improved to almost double for an even number  $N$ . The following theorem is necessary for constructing a more practical algorithm:

**Theorem 3.2.11.** *Let the Gegenbauer polynomials be standardized by Equation (3.A.2). Then for even values of  $N$ , we have*

$$\int_{-1}^a C_{N+1}^{(\alpha)}(x)dx = \int_{-1}^{-a} C_{N+1}^{(\alpha)}(x)dx \quad \forall \alpha > -\frac{1}{2}; a \in [-1, 1]. \quad (3.36)$$

*Proof.* The proof can be easily verified using the symmetry property (3.A.1), and the Gegenbauer integration formulas (3.A.11).  $\square$

The symmetry of the numerator  $\int_{-1}^{x_i} C_{M+1}^{(\alpha_i^*)}(x)dx$  in  $\eta_{i,M}(\alpha_i^*)$  permits the reduction of most of the calculations encountered in Steps 4-6 in Algorithm 2.1 by half. In fact, since we have  $\alpha_i^* = \alpha_{N-i}^* \forall i = 0, \dots, N/2 - 1$ , then the adjoint GG points  $z_{i,j} \in S_{N,M}$  can be stored for the first  $N/2 - 1$  iterations, and invoked later in the next iterations. The same can be carried out for the values of the Gegenbauer polynomials  $C_j^{(\alpha_i^*)}(z_{i,m})$ , and the parameters  $\lambda_j^{(\alpha_i^*)}; \omega_j^{(\alpha_i^*)} \forall 0 \leq j, m \leq M$ . Hence for symmetric sets of integration points  $\{x_i\}_{i=0}^N$ , Algorithm 2.1 can be reduced to a more cost-effective algorithm, Algorithm 2.2 (see Appendix 3.H).

The reader should notice that Algorithm 2.1 is suitable for general sets of integration nodes distributed arbitrarily along the interval  $[-1, 1]$ , while Algorithm 2.2 is valid only for arbitrary symmetric sets of integration nodes along the interval  $[-1, 1]$ . We notice that Algorithms 2.1 & 2.2 are flexible enough to construct the suitable quadrature in the sought error norm. For instance, if the approximations are sought in the  $L^\infty$ -norm, then Chebyshev polynomial expansions are preferable to Legendre polynomial expansions as discussed before. Therefore Step 1 in both algorithms can be implemented using the Chebyshev matrix. Moreover, Steps 5 & 6 in Algorithms 2.1 & 2.2, respectively, can be implemented by choosing  $\alpha_i^* \in \{-0.5 + \varepsilon, 0\}$ . Although the PMQ converges automatically to the Chebyshev quadrature for large values of  $M$ , omitting Step 1 in both algorithms (implemented using the Chebyshev matrix) is not recommended, as we already know the exact optimal value of  $\alpha_i^*$  for large values of  $M$  in this

## Chapter 3

---

case; moreover, the line search method may not determine the exact value  $\alpha_i^* = 0$  of the optimal Gegenbauer parameter, due to the round-off errors encountered during the calculations. In the following section we shall demonstrate the robustness and the advantages of the developed PMQ, via some extensive numerical experiments.

### 3.3 Numerical Results

In this section, we conduct many test experiments to show the higher accuracy and the faster convergence rates achieved by the PMQ over the optimal QMQ, and the Clenshaw-Curtis (CC) method. The numerical results of the optimal QMQ are as quoted from Ref. (El-Hawary et al., 2000). The numerical results of the P-matrix were obtained via Algorithm 2.2 with  $M_{\max} = 12$ . All numerical experiments were conducted on a personal laptop with a 2.53 GHz Intel Core i5 CPU and 4G memory running on a Windows 7 operating system using a FORTRAN compiler in double-precision real arithmetic. The domain of the integration points  $\{x_i\}_{i=0}^N$  considered in this section is transformed into the interval  $[0, 1]$  using the change of variable  $t = (1 + x)/2, x \in [-1, 1]$ , unless stated otherwise. The Elgindy and Hedar (2008) line search method was implemented to determine the locally optimal Gegenbauer parameters  $\{\alpha_i^*\}_{i=0}^N$ . The interval  $[-0.5 + 2\tilde{\varepsilon}, 2]$  was chosen as the initial uncertainty interval in the line search method, where  $\tilde{\varepsilon} = 2.22 \times 10^{-16}$  is the machine epsilon, and the right endpoint 2 is the maximum plausible Gegenbauer parameter as discussed earlier in Section 3.2.6. The line search method was stopped if the approximate solution satisfied the following stopping criteria:

$$\left| \frac{d}{d\alpha}(\eta_{i,M}^2(\alpha)) \right| < 10^{-16} \wedge \frac{d^2}{d\alpha^2}(\eta_{i,M}^2(\alpha)) > 0 \quad \forall 0 \leq i \leq N.$$

#### 3.3.1 Comparisons with the Optimal QMQ

In this section we are interested in comparing the accuracy of the PMQ versus the optimal QMQ. Three test problems with reported data and results have been quoted from Ref. (El-Hawary et al., 2000). The numerical test functions are generally smooth functions studied earlier by Don and Solomonoff (1997) and Gustafson and Silva (1998). The integrations of the following three test functions over the interval  $[0, 1]$  are considered:

$$f_1(x) = e^{2x}, f_2(x) = \sin(2x); f_3(x) = \frac{1}{\sqrt{1+x}}.$$

## Chapter 3

---

The numerical experiments conducted in (El-Hawary et al., 2000) show that the optimal QMQ obtained through the solution of Problem (3.8) produces higher-order approximations better than those obtained by the Chebyshev and Legendre quadratures for some examples. Hence it is sufficient to show that the PMQ outperforms the optimal QMQ in accuracy through the same test functions to demonstrate that the PMQ can produce better approximations than both the standard Chebyshev and Legendre quadratures for small-scale number of spectral expansion terms. Tables 3.1-3.3 show the maximum absolute errors (MAEs) for the three test functions  $\{f_i\}_{i=1}^3$  obtained using a square P-matrix of size  $(N + 1)$  for  $N = 4, 6, 8, 10; 12$ , and different sets of integration nodes. Notice here that Algorithm 2.2 skips its first step, and seeks to establish an optimal Gegenbauer quadrature for each integration point  $\{x_i\}_{i=0}^N$ . The results in the tables are listed as follows: The first column  $N$  denotes the highest degree of the Gegenbauer polynomial approximation of the definite integrals of the test functions  $\{f_i\}_{i=1}^3$ . The second column gives the results obtained by El-Hawary et al. (2000) in the form  $\alpha^*/\text{MAE}$ , where  $\alpha^*$  denotes the optimal Gegenbauer parameter obtained by solving Problem (3.8). The last four columns show the MAE obtained using the PMQ for the following four sets of integration points: The sets  $S_N^{(0)}$ ,  $S_N^{(\alpha^*)}$ , the set  $S^{3,N}$  of Chebyshev-Gauss-Lobatto (CGL) points  $x_i = -\cos(i\pi/N)$ ,  $i = 0, \dots, N$ ; the set  $S^{4,N}$  of equispaced points  $x_i = -1 + 2i/N$ ,  $i = 0, \dots, N$ . The experiments in Tables 3.1-3.3 were implemented using the values of  $\varepsilon = 0.016, 0.028; 0.01$ , respectively.

Table 3.1: The PMQ versus the optimal QMQ in approximating the definite integrals of  $f_1(x)$  for the set of integration nodes  $S_N^{(\alpha^*)}$ .

$N$	Optimal QMQ (El-Hawary et al., 2000)	Present PMQ			
	$\alpha^*/\text{MAE}$	$S_N^{(0)}$	$S_N^{(\alpha^*)}$	$S^{3,N}$	$S^{4,N}$
4	0.5080/ $1.842 \times 10^{-04}$	$9.212 \times 10^{-05}$	$9.212 \times 10^{-05}$	$9.212 \times 10^{-05}$	$9.212 \times 10^{-05}$
6	0.4982/ $8.100 \times 10^{-07}$	$3.950 \times 10^{-07}$	$4.064 \times 10^{-07}$	$2.851 \times 10^{-07}$	$3.240 \times 10^{-07}$
8	0.5278/ $2.196 \times 10^{-09}$	$1.099 \times 10^{-09}$	$1.099 \times 10^{-09}$	$1.099 \times 10^{-09}$	$1.099 \times 10^{-09}$
10	0.5323/ $4.091 \times 10^{-12}$	$2.045 \times 10^{-12}$	$2.069 \times 10^{-12}$	$1.836 \times 10^{-12}$	$1.359 \times 10^{-12}$
12	3.2000/ $1.055 \times 10^{-13}$	$2.442 \times 10^{-15}$	$9.104 \times 10^{-15}$	$2.220 \times 10^{-15}$	$2.331 \times 10^{-15}$

Table 3.2: The PMQ versus the optimal QMQ in approximating the definite integrals of  $f_2(x)$  for the set of integration nodes  $S_N^{(\alpha^*)}$ .

$N$	Optimal QMQ (El-Hawary et al., 2000)	Present PMQ			
	$\alpha^*/\text{MAE}$	$S_N^{(0)}$	$S_N^{(\alpha^*)}$	$S^{3,N}$	$S^{4,N}$
4	0.5191/ $1.758 \times 10^{-05}$	$1.695 \times 10^{-05}$	$1.695 \times 10^{-05}$	$1.695 \times 10^{-05}$	$1.695 \times 10^{-05}$
6	0.4854/ $7.854 \times 10^{-08}$	$7.611 \times 10^{-08}$	$7.756 \times 10^{-08}$	$5.553 \times 10^{-08}$	$5.941 \times 10^{-08}$
8	0.5110/ $2.110 \times 10^{-10}$	$2.098 \times 10^{-10}$	$2.098 \times 10^{-10}$	$2.098 \times 10^{-10}$	$2.098 \times 10^{-10}$
10	0.4800/ $4.021 \times 10^{-13}$	$3.983 \times 10^{-13}$	$4.004 \times 10^{-13}$	$3.583 \times 10^{-13}$	$2.572 \times 10^{-13}$
12	3.2000/ $8.326 \times 10^{-15}$	$5.274 \times 10^{-16}$	$4.441 \times 10^{-16}$	$4.718 \times 10^{-16}$	$4.718 \times 10^{-16}$

## Chapter 3

Table 3.3: The PMQ versus the optimal QMQ in approximating the definite integrals of  $f_3(x)$  for the set of integration nodes  $S_N^{(\alpha^*)}$ .

$N$	Optimal QMQ (El-Hawary et al., 2000)	Present PMQ			
	$\alpha^*/\text{MAE}$	$S_N^{(0)}$	$S_N^{(\alpha^*)}$	$S^{3,N}$	$S^{4,N}$
4	0.5350/3.729 $\times 10^{-06}$	3.637 $\times 10^{-06}$	3.451 $\times 10^{-06}$	3.451 $\times 10^{-06}$	3.451 $\times 10^{-06}$
6	0.4918/7.344 $\times 10^{-08}$	7.174 $\times 10^{-08}$	7.146 $\times 10^{-08}$	5.082 $\times 10^{-08}$	5.363 $\times 10^{-08}$
8	0.4910/1.525 $\times 10^{-09}$	1.435 $\times 10^{-09}$	1.462 $\times 10^{-09}$	1.357 $\times 10^{-09}$	1.357 $\times 10^{-09}$
10	0.5110/3.327 $\times 10^{-11}$	3.241 $\times 10^{-11}$	3.252 $\times 10^{-11}$	2.903 $\times 10^{-11}$	2.025 $\times 10^{-11}$
12	0.5118/7.765 $\times 10^{-13}$	7.562 $\times 10^{-13}$	7.749 $\times 10^{-13}$	7.042 $\times 10^{-13}$	7.042 $\times 10^{-13}$

Tables 3.1-3.3 show that the PMQ outperforms the optimal QMQ for all three test functions  $\{f_i\}_{i=1}^3$ , and all sets of integration points  $\{x_i\}_{i=0}^N$  considered. In Table 3.1, the comparison between the PMQ and the optimal QMQ shows an advantage of the PMQ of one decimal figure of accuracy at  $N = 4$ . The difference then reaches two decimal figures of accuracy at  $N = 12$ . In Table 3.2, the two quadratures start with the same order of accuracy with slight improvements in precision in favor of the PMQ. The PMQ then achieves almost full machine precision at  $N = 12$  for all four sets of integration points, with faster convergence rate than the optimal QMQ. The two quadratures achieve almost the same convergence rates for the rational function  $f_3(x)$ , with slight improvements in accuracy in favor of the PMQ. Figure 3.3 shows the values of  $\alpha_i^*$  versus the four sets of integration points  $S_{12}^{(0)}, S_{12}^{(\alpha^*)}, S^{3,12}, S^{4,12}$ . It can be clearly seen from the figure that Algorithm 2.2 can employ some better choices of the Gegenbauer polynomials  $C_M^{(\alpha_i^*)}(x)$  than Chebyshev and Legendre polynomials, mitigating the effect of the quadrature error throughout the range of the integration.

We notice that faster rates of convergence of the PMQ are achieved for the exponential and the trigonometric test functions  $f_1(x); f_2(x)$ , respectively, while slower convergence rates are observed for the rational test function  $f_3(x)$ . In fact, the construction of the PMQ via the pointwise approach presented in Section 3.2 is induced by the type of the integration points set  $\{x_i\}_{i=0}^N$  regardless of the integrand function  $f$ . Nonetheless, the magnitude of the quadrature error depends on the magnitude of the  $(N + 1)$ th derivative of  $f$ . In fact,  $\max_{0 \leq x \leq 1} |f^{(N+1)}(x)|$  for the three functions  $\{f_i\}_{i=0}^3$  is a monotonically increasing function for increasing values of  $N$ , which exhibits its slowest and fastest increase rates for the trigonometric function  $f_2(x)$  and the rational function  $f_3(x)$ , respectively. In particular, at the value  $N = 12$ ,  $\max_{0 \leq x \leq 1} |f_2^{(N+1)}(x)| = 8192$ , while  $\max_{0 \leq x \leq 1} |f_3^{(N+1)}(x)| \approx 9.65 \times 10^8$ , which explains the clear discrepancies in the convergence rates of the PMQ for the three test functions. Moreover, we notice that the rational test function  $f_3(x)$  is analytic in the neighborhood of  $[-1, 1]$ , but not throughout the complex plane; thus we see that the convergence rates of the

## Chapter 3

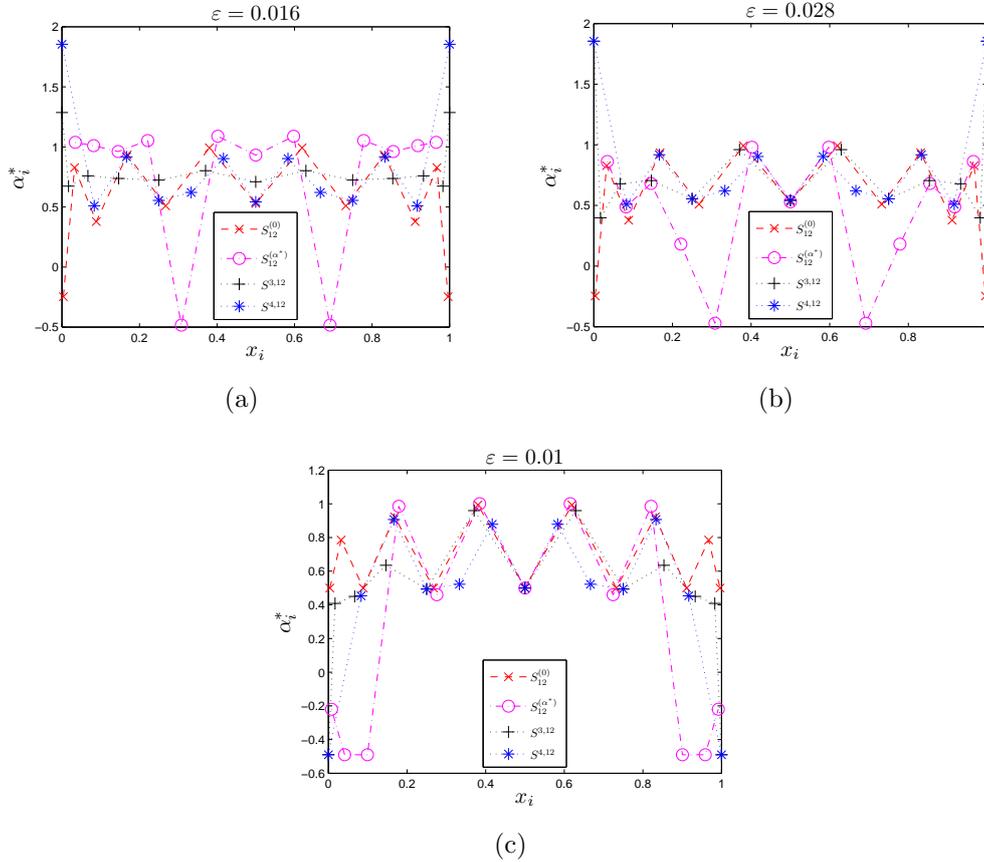


Figure 3.3: The values of  $\alpha_i^*$  versus the four sets of integration points  $S_{12}^{(0)}$ ,  $S_{12}^{(\alpha^*)}$ ,  $S^{3,12}$ ;  $S^{4,12}$  using  $\varepsilon = 0.016, 0.028; 0.01$ .

two quadrature methods are quite close to each other for the last test example.

The results in Tables 3.1-3.3 seem to suggest that the equispaced nodes  $x_i \in S^{4,N}$  are just about as good as point sets typical of Gaussian quadrature algorithms in terms of carrying out the numerical quadrature. This in turn seems to contradict the fact that numerical quadratures using polynomial interpolants of high degrees, such as the global Newton-Cotes methods of increasing orders, go exponentially unstable except for special cases as  $N$  is increased for interpolation based on equally spaced nodes (a problem widely known as the “Runge Phenomenon”). In fact, although the PMQ (3.14) can use the equally spaced points as the integration points set, the construction of the numerical quadrature scheme is always carried out through interpolations at the adjoint GG points  $z_{i,j} \in S_{N,M}$ , which minimize the quadrature truncation error in the sense of solving Problem (3.13). This efficient numerical scheme is established for any arbitrary sets of

## Chapter 3

---

integration points. Hence the PMQ can stabilize the calculations for any arbitrary sets of integration points, while avoiding the Runge Phenomenon. Notice that the GG points are exactly those which minimize the maximum value of the polynomial factor of the interpolation error (Kopriva, 2009), and the convergence is guaranteed if the function is analytic on  $[-1, 1]$  (Trefethen, 1996). On the other hand, the optimal QMQ approximations are only obtained for the GG integration points set  $S_N^{(\alpha^*)}$ , since the discretizations are carried out using the same set of GG points  $S_N^{(\alpha^*)}$ .

The numerical results of Tables 3.1-3.3 clearly refute the general claim that the optimal QMQ ‘gives an optimal approximation of the integrals’ (El-Hawary et al., 2000). Moreover, Tables 3.1-3.3 show that better integration approximations can be sought for arbitrary sets of integration points  $\{x_i\}_{i=0}^N$  rather than the limited choice of the GG integration points set  $S_N^{(\alpha)}$ . These numerical experiments support the theoretical arguments presented in Section 3.2, which demonstrate that breaking up the minimax problem (3.10) into several subproblems via a pointwise approach can produce higher-order approximations of integrals.

Perhaps one of the most important features of the P-matrix appears in its rectangular form, which permits faster convergence rates by increasing the number  $(M + 1)$  of the P-matrix columns. We stress here that the number  $M$  is independent of the number of integration points  $(N + 1)$ , and its value controls the number of function/quadrature evaluations. Table 3.4 shows the MAE obtained by the PMQ versus the optimal QMQ for the same number  $(N + 1)$  of integration nodes, and increasing values of  $M$ . The results of the PMQ are reported for the test function  $f_1(x)$  and the integration points set  $S_N^{(\alpha^*)}$  obtained in (El-Hawary et al., 2000). The first column of the table lists the values of  $N$ . The second column lists the values of  $\alpha^*$  and the MAE in the form  $\alpha^*/\text{MAE}$ . The remaining columns list the values of  $M$  and the MAE in the form  $M/\text{MAE}$ . The table shows that the order of the PMQ is not coupled with the value of  $N$ , where we can clearly see that increasing the number  $(M + 1)$  of the P-matrix columns increases the order of the approximation, for each value of  $N$ . This key result is a major element in increasing the order of the Gegenbauer collocation integration schemes without the requirement of increasing the number of collocation points; cf. (Elgindy and Smith-Miles, 2013c).

## Chapter 3

---

Table 3.4: The PMQ versus the optimal QMQ in approximating the definite integrals of  $f_1(x)$  for the set of integration nodes  $S_N^{(\alpha^*)}$ . The results of the PMQ are reported for increasing values of  $M$ .

$N$	Optimal QMQ (El-Hawary et al., 2000)	Present PMQ				
	$\alpha^*/\text{MAE}$	$M/\text{MAE}$	$M/\text{MAE}$	$M/\text{MAE}$	$M/\text{MAE}$	$M/\text{MAE}$
4	0.5080/1.842 $\times 10^{-04}$	4/9.212 $\times 10^{-05}$	6/3.029 $\times 10^{-08}$	8/1.099 $\times 10^{-09}$	10/1.089 $\times 10^{-13}$	12/2.220 $\times 10^{-15}$
6	0.4982/8.100 $\times 10^{-07}$			8/1.099 $\times 10^{-09}$	10/1.130 $\times 10^{-12}$	12/7.994 $\times 10^{-15}$
8	0.5278/2.196 $\times 10^{-09}$			8/1.099 $\times 10^{-09}$	10/1.328 $\times 10^{-12}$	12/2.220 $\times 10^{-15}$
10	0.5323/4.091 $\times 10^{-12}$			10/2.045 $\times 10^{-12}$	12/2.331 $\times 10^{-15}$	

In fact, the rectangular form of the P-matrix is not only useful for increasing the order of the approximation if the number of integration points is relatively small, but also can be very effective in limiting the effect of the round-off errors associated with the approximation of the definite integrals by a square spectral integration matrix of a large-scale size ( $N + 1$ ). In particular, although spectral integration matrices are much more stable operators than spectral differentiation matrices, the effect of the round-off error may arise from the large number of matrix-vector operations required by a typical square spectral integration matrix of a large-scale size. For instance, consider the problem of approximating the definite integrals  $\int_{-1}^{x_i} f(x)dx$  of the function  $f(x) = (1 - x^2)e^x$ , for the CGL points  $\{x_i\}_{i=0}^{256}$ . This problem was studied by Elgindy (2009) using a square Chebyshev pseudospectral integration matrix. The MAEs reported using Equations (4.6) & (4.7) in (Elgindy, 2009) were  $1.99840 \times 10^{-15}$ ;  $2.22045 \times 10^{-15}$ , respectively. Approximating the same integrals using the PMQ with  $M = 16$  (without transforming the integration domain into  $[0, 1]$ ) establishes a more stable numerical scheme with the MAE of  $6.66134 \times 10^{-16}$ . Hence the rectangular form of the P-matrix allows one to seek high-order approximations for a large number ( $N + 1$ ) of integration points using a relatively small number ( $M + 1$ ) of Gegenbauer expansion terms. In contrast, the size of the conventional spectral integration matrices based on discretizations at the very same integration points set (usually of the Gauss, or Gauss-Lobatto points type) must be the same as the value of ( $N + 1$ ).

### 3.3.2 Comparisons with the CC Method

In this section we conduct further numerical experiments on the PMQ to test its accuracy versus the popular CC method originally developed by Clenshaw and Curtis in 1960; cf. (Clenshaw and Curtis, 1960). The latter method is based on the expansion of the integrand functions in terms of the Chebyshev polynomials. Here we applied the CC method using Algorithms CHEBFT, CHINT; CHEBEV given in (Press et al., 1992). The numerical experiments were conducted on the three test functions  $\{f_i\}_{i=1}^3$  presented in the previous section. Moreover, we have extended the numerical experiments to include the following more challenging

## Chapter 3

---

test functions:

$$f_4(x) = x^{20}, f_5(x) = e^{-x^2}, f_6(x) = \frac{1}{1 + 25x^2}; f_7(x) = e^{-1/x^2}.$$

The test function  $f_4(x)$  is a monic polynomial. The test function  $f_5(x)$  is entire, i.e., analytic throughout the complex plane. Both functions  $f_4; f_5$  exhibit strong growths  $O(|x|^{20}); O(e^{|x|^2})$  as  $x \rightarrow \infty$  in the complex plane. The test function  $f_6(x)$  is analytic in a neighborhood of  $[-1, 1]$ , but not throughout the complex plane. This test function is known to be a troublesome for high-degree polynomial interpolants at equally spaced nodes (Kopriva, 2009; Mason and Handscomb, 2003). The last test function  $f_7(x)$  is a smooth function throughout the complex plane. The reported results of the CC method are all obtained at the zeros of the Chebyshev polynomials  $x_i \in S_N^{(0)}$ . The results of the PMQ were obtained for the three sets of integration points  $S_N^{(0)}, S^{3,N}; S^{4,N}$  using a square P-matrix of size  $(N + 1)$ . Moreover, the results of the PMQ for the test functions  $\{f_i(x)\}_{i=4}^7$  were obtained by setting  $\alpha_i^* = 0.5, i = 0, \dots, N$ , for all critical  $\alpha$ -values determined from Step 5 in Algorithm 2.2.

Figure 4 shows the logarithm of the Euclidean error (EE) of the PMQ and the CC quadrature (CCQ) versus  $N = 2, 4, \dots, 20$ , for the seven test functions  $\{f_i\}_{i=1}^7$  on  $[0, 1]$ . Figs. 3.4(a) and 3.4(b) show the numerical experiments conducted on the first two test functions  $f_1(x) = e^{2x}; f_2(x) = \sin(2x)$ . It can be clearly seen from the figures that the PMQ manifests faster convergence rates than CCQ for a small scale of the number of spectral expansion terms  $N$ , in particular, in the range  $1 \leq N \leq 12$ . After this stage the two methods share almost the same convergence rates, as the precisions of the resulting approximations nearly reach full machine precision. We notice that this convergence behavior is practically the same for all sets of integration points. Figure 3.4(c) shows the numerical results obtained for the third test function  $f_3(x) = 1/\sqrt{1+x^2}$ , where the rapid convergence of the PMQ is clearly observed in the range  $1 \leq N \leq 16$ . The two methods then tend to produce nearly the same orders of accuracy for increasing values of  $N$ . Figure 3.4(d) reports the numerical results for the fourth test function  $f_4(x) = x^{20}$ . Here we notice that the PMQ converges much faster than the CCQ over the whole range  $1 \leq N \leq 20$ . This suggests that the PMQ is very effective for problems with strongly growing solutions. Figure 3.4(e) shows the numerical results obtained for the fifth test function  $f_5(x) = e^{-x^2}$ . The rapid convergence rate of the PMQ is conspicuous for the values of  $1 \leq N \leq 16$ , where a 14<sup>th</sup>-order Legendre polynomial expansion is sufficient to achieve nearly full machine precision. The two methods then share similar convergence rates. We notice that the behavior of the PMQ is almost the same for all sets of integration points. Figs. 3.4(f) and 3.4(g) report the numerical results for the sixth and the seventh test functions  $f_6(x) = 1/(1 + 25x^2); f_7(x) = e^{-1/x^2}$ . Here we notice

## Chapter 3

---

that the convergence rates of the PMQ for both test functions exceed that of the CCQ over the whole range  $1 \leq N \leq 20$ . We notice also that for these two test functions, the behaviors of the PMQ can be distinguished for the three sets of integration points, where the choice of  $S^{3,N}$  produces the best approximations for most values of  $N$ . All of the above results confirm, without doubt, that other members of the parent family of Gegenbauer polynomials converge faster than the Chebyshev polynomials for the small range of the spectral expansion series; moreover, the PMQ provides a means for determining these effective members.

**Remark 3.3.1.** *The optimal Gegenbauer parameters obtained in Section 3.3 were determined locally using the Elgindy and Hedar (2008) line search method. However, any global line search method can be applied to determine the global optimal Gegenbauer parameters.*

### 3.4 Further Applications

Spectral methods are now a popular tool for the solution of ordinary and partial differential equations, integral and integro-differential equations, etc.; cf. (Canuto et al., 2006, 2007; Driscoll, 2010; Elgindy, 2009; Elgindy and Smith-Miles, 2013c; Hesthaven et al., 2007; Kopriva, 2009; Tang et al., 2008). In (Elgindy and Smith-Miles, 2013c), we applied the PMQ together with the standard  $\hat{\text{P}}\text{MQ}$  for solving BVPs, integral, and integro-differential equations in the physical space. Our work in these areas established an efficient Gegenbauer collocation numerical method, which generally leads to well-conditioned linear systems, and avoids the degradation of precision caused by severely ill-conditioned spectral differentiation matrices. Perhaps one of the most significant contributions of this recent work is the ability of the Gegenbauer collocation integration method to achieve spectral accuracy using a very small number of solution points—almost full machine precision in some cases, which is highly desirable in these areas. The numerical experiments conducted show that the Gegenbauer polynomial expansions associated with their optimal integration matrices can produce very accurate approximations, which have precision exceeding that obtained by the standard Chebyshev, Legendre, and Gegenbauer polynomial expansions for the small/medium range of the number of spectral expansion terms.

Optimal control theory represents another vital area in mathematics where numerical integration can play a great role. Indeed, a closed form expression for the optimal control is usually out of reach, and classical solution tools such as the calculus of variations, dynamic programming, and Pontryagin's maximum/minimum principle can only provide the analytical optimal control in very special cases. Therefore, numerical methods for solving optimal control problems

## Chapter 3

---

are extremely important. There are several classes of methods for solving optimal control problems, but one of the most promising numerical methods which came into prominence in the past three decades is the class of direct orthogonal collocation methods (Benson, 2004; Benson et al., 2006; Elgindy et al., 2012; Elnagar, 1997; Elnagar and Kazemi, 1995; Elnagar and Razzaghi, 1997; Fahroo and Ross, 2008; Garg et al., 2011b; Vlassenbroeck and Dooren, 1988). The aim of these methods is to transcribe the infinite-dimensional continuous-time optimal control problem into a finite-dimensional parameter nonlinear programming problem via discretization of the original problem in time, and performing some parameterizations of the control and/or state vectors. A full parameterization is typically carried out by expanding the states and the controls by using the spectral expansion series in terms of some prescribed global basis polynomials, frequently chosen as the Chebyshev and Legendre polynomials. For problems with sufficiently differentiable states and controls, the PMQ can produce very precise approximations to the integrals involved in the components of the optimal control problem. Our work in (Elgindy et al., 2012) shows that the higher-order approximations produced by the PMQ can significantly reduce the number of terms in the Gegenbauer expansion series approximating the states and the controls; consequently, the dimension of the resulting nonlinear programming problem is significantly reduced, and it can be solved readily using standard nonlinear programming software. Moreover, since the P-matrix is independent of the integrand function, it is a constant matrix for a particular set of integration nodes, and can be used to solve practical trajectory optimization problems quickly.

### 3.5 Future Work

Most of the interesting questions concerning the application of the PMQ in solving general mathematical problems remain open. Perhaps one of the main topics to be pursued later occurs in the development of high-order numerical quadratures for approximating the integrals of general nonsmooth functions. In fact, it is well-known that the local discontinuities of a function ruin the convergence of the global spectral approximations ‘even in regions for which the underlying function is analytic’ (Gottlieb et al., 2011). One approach for overcoming this problem is to apply the Gegenbauer reconstruction method to eliminate the Gibbs phenomenon from the spectral approximation while maintaining its exponential convergence properties, even up to the discontinuities of the function (Gelb and Jackiewicz, 2005). The interested readers may consult Refs. (Gelb and Gottlieb, 2007; Gelb and Jackiewicz, 2005; Gottlieb et al., 2011; Gustafsson, 2011; Jackiewicz and Park, 2009; Jung et al., 2010) for an overview on this significant contribution, and some of the developments achieved in this trend.

### 3.6 Conclusion

This chapter reports a novel optimal Gegenbauer quadrature method for approximating definite integrations. The proposed method gathers the useful properties and the main strengths of the Chebyshev, Legendre, and Gegenbauer polynomials in one optimal numerical quadrature. The unified numerical scheme developed in this chapter shows that the Gegenbauer polynomial expansion series  $\sum_{k=0}^N a_k C_k^{(\alpha)}(x)$  can produce higher-order approximations to the integrals  $\int_{-1}^{x_i} f(x)dx$  of some given function  $f(x) \in C^\infty[-1, 1]$  for the small range of the expansion terms by minimizing the quadrature error at each integration point  $x_i$ . This technique entails the calculation of some optimal Gegenbauer parameter values  $\alpha_i^*$  rather than choosing any arbitrary  $\alpha$  value. For a large-scale number of expansion terms, the PMQ provides the advantage of convergence to the optimal Chebyshev and Legendre quadratures in the  $L^\infty$ -norm and  $L^2$ -norm, respectively. The developed computational Algorithms 2.1 & 2.2 construct the PMQ through interpolations at an optimal set of adjoint GG points in the sense of solving Problem (3.13). Algorithm 2.2 is a more cost-effective algorithm suitable for similar sets of integration points, where most of the calculations carried out for the construction of the P-matrix are halved; thus the Gegenbauer spectral computations can be considerably more effective. The construction of the developed PMQ is induced by the set of integration points regardless of the integrand function. The proposed method establishes a high-order numerical quadrature for any arbitrary sets of integration points, and avoids the Runge Phenomenon through discretizations at the adjoint GG points. The rectangular form of the developed P-matrix permits rapid convergence rates without the need to increase the number of integration nodes. The PMQ is exact for polynomials of any arbitrary degree  $n$  if the number of columns of the P-matrix is greater than or equal to  $n$ . Our proposed method is strong enough to stabilize the calculations, and sufficient to retain the spectral accuracy. The numerical experiments reported in this chapter show that the PMQ can achieve very rapid convergence rates, and higher-order precisions, which can exceed those obtained by the standard Chebyshev and Legendre polynomial methods. Moreover, the PMQ outperforms conventional Gegenbauer quadrature methods. The present method can be applied for solving mathematical problems of several types including integral equations, integro-differential equations, BVPs, and optimal control problems.

## Chapter 3

---

### 3.A Some Properties of the Gegenbauer Polynomials

The Gegenbauer polynomial  $C_n^{(\alpha)}(x)$  of degree  $n$  and associated with the parameter  $\alpha > -1/2$  is a real-valued function. It appears as an eigensolution to the following singular Sturm-Liouville problem in the finite domain  $[-1, 1]$  (Szegő, 1975):

$$\frac{d}{dx}(1-x^2)^{\alpha+\frac{1}{2}} \frac{dC_n^{(\alpha)}(x)}{dx} + n(n+2\alpha)(1-x^2)^{\alpha-\frac{1}{2}} C_n^{(\alpha)}(x) = 0.$$

The weight function for the Gegenbauer polynomials is the even function  $w^{(\alpha)}(x) = (1-x^2)^{\alpha-1/2}$ , and the orthogonality relation of the Gegenbauer polynomials standardized by Szegő (1975) is given by

$$\int_{-1}^1 (1-x^2)^{\alpha-\frac{1}{2}} C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = h_n^{(\alpha)} \delta_{mn},$$

where

$$h_n^{(\alpha)} = \frac{2^{1-2\alpha} \pi \Gamma(n+2\alpha)}{n!(n+\alpha)\Gamma^2(\alpha)},$$

is the normalization factor;  $\delta_{mn}$  is the Kronecker delta function. The symmetry of the Gegenbauer polynomials is emphasized by the relation (Hesthaven et al., 2007)

$$C_n^{(\alpha)}(x) = (-1)^n C_n^{(\alpha)}(-x). \quad (3.A.1)$$

Doha (1990) standardized the Gegenbauer polynomials so that

$$C_n^{(\alpha)}(1) = 1, \quad n = 0, 1, 2, \dots \quad (3.A.2)$$

This standardization establishes the useful relations that  $C_n^{(0)}(x)$  becomes identical with the Chebyshev polynomial of the first kind  $T_n(x)$ ,  $C_n^{(1/2)}(x)$  is the Legendre polynomial  $L_n(x)$ ; and  $C_n^{(1)}(x)$  is equal to  $(1/(n+1))U_n(x)$ , where  $U_n(x)$  is the Chebyshev polynomial of the second type. Throughout the chapter, when we refer to the Gegenbauer polynomials, we mean those standardized by Equation (3.A.2). Also when we refer to the Chebyshev polynomials, we mean the Chebyshev polynomials of the first kind. Using Standardization (3.A.2), the Gegenbauer polynomials are generated by using Rodrigues' formula in the following form:

$$C_n^{(\alpha)}(x) = \left(-\frac{1}{2}\right)^n \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} (1-x^2)^{\frac{1}{2}-\alpha} \frac{d^n}{dx^n} ((1-x^2)^{n+\alpha-\frac{1}{2}}), \quad (3.A.3)$$

## Chapter 3

---

or starting with the following two equations:

$$C_0^{(\alpha)}(x) = 1, \quad (3.A.4a)$$

$$C_1^{(\alpha)}(x) = x, \quad (3.A.4b)$$

the Gegenbauer polynomials can be generated directly by using the following three-term recurrence equation:

$$(j + 2\alpha)C_{j+1}^{(\alpha)}(x) = 2(j + \alpha)x C_j^{(\alpha)}(x) - j C_{j-1}^{(\alpha)}(x), \quad j \geq 1. \quad (3.A.4c)$$

Using Standardization (3.A.2) and Equation (4.7.1) in (Szegő, 1975), one can readily show that the Gegenbauer polynomials  $C_n^{(\alpha)}(x)$  and the Gegenbauer polynomials  $\hat{C}_n^{(\alpha)}(x)$  standardized by Szegő (1975) are related by

$$C_n^{(\alpha)}(x) = \frac{\hat{C}_n^{(\alpha)}(x)}{\hat{C}_n^{(\alpha)}(1)} \quad \forall x \in [-1, 1], \alpha > -\frac{1}{2}; n \geq 0. \quad (3.A.5)$$

Hence the Gegenbauer polynomials  $C_n^{(\alpha)}(x)$  satisfy the orthogonality relation

$$\int_{-1}^1 (1 - x^2)^{\alpha - \frac{1}{2}} C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = \lambda_n^{(\alpha)} \delta_{mn}, \quad (3.A.6)$$

where

$$\lambda_n^{(\alpha)} = \frac{2^{2\alpha-1} n! \Gamma^2(\alpha + \frac{1}{2})}{(n + \alpha) \Gamma(n + 2\alpha)}, \quad (3.A.7)$$

is the normalization factor;  $\delta_{mn}$  is the Kronecker delta function. Moreover, the leading coefficients  $K_j^{(\alpha)}$  of the Gegenbauer polynomials  $C_j^{(\alpha)}(x)$  are

$$K_j^{(\alpha)} = 2^{j-1} \frac{\Gamma(j + \alpha) \Gamma(2\alpha + 1)}{\Gamma(j + 2\alpha) \Gamma(\alpha + 1)}, \quad (3.A.8)$$

for each  $j$ . The orthonormal Gegenbauer basis polynomials are defined by

$$\phi_j^{(\alpha)}(x) = (\lambda_j^{(\alpha)})^{-\frac{1}{2}} C_j^{(\alpha)}(x), \quad j = 0, \dots, n, \quad (3.A.9)$$

and they satisfy the discrete orthonormality relation

$$\sum_{j=0}^n \omega_j^{(\alpha)} \phi_s^{(\alpha)}(x_j) \phi_k^{(\alpha)}(x_j) = \delta_{sk}, \quad (3.A.10)$$

## Chapter 3

---

where the  $x_j \in S_n^{(\alpha)}$ ;  $\omega_j^{(\alpha)}$  are as defined by Equations (3.3) & (3.6), respectively. The integrations of the Gegenbauer polynomials  $C_j^{(\alpha)}(x)$  can be obtained through Equations (3.A.4) as follows (El-Hawary et al., 2000):

$$\int_{-1}^x C_0^{(\alpha)}(x)dx = C_0^{(\alpha)}(x) + C_1^{(\alpha)}(x), \quad (3.A.11a)$$

$$\int_{-1}^x C_1^{(\alpha)}(x)dx = a_1(C_2^{(\alpha)}(x) - C_0^{(\alpha)}(x)), \quad (3.A.11b)$$

$$\int_{-1}^x C_j^{(\alpha)}(x)dx = \frac{1}{2(j+\alpha)}(a_2 C_{j+1}^{(\alpha)}(x) + a_3 C_{j-1}^{(\alpha)}(x) + (-1)^j(a_2 + a_3)), \quad j \geq 2, \quad (3.A.11c)$$

where

$$a_1 = \frac{1+2\alpha}{4(1+\alpha)}, \quad a_2 = \frac{j+2\alpha}{(j+1)}, \quad a_3 = -\frac{j}{(j+2\alpha-1)}.$$

For further information about the Gegenbauer polynomials, the interested reader may consult (Abramowitz and Stegun, 1965; Bayin, 2006; Szegö, 1975).

### 3.B Proof of Theorem 3.2.3

Since  $f(z_{i,j}) = \sum_{k=0}^M a_{i,k} C_k^{(\alpha_i^*)}(z_{i,j})$ ,  $i = 0, \dots, N$ ;  $j = 0, \dots, M$ , for some Gegenbauer coefficients  $a_{i,k}$ , then

$$\begin{aligned} \omega_j^{(\alpha_i^*)} C_s^{(\alpha_i^*)}(z_{i,j}) f(z_{i,j}) &= \sum_{k=0}^M a_{i,k} \omega_j^{(\alpha_i^*)} C_s^{(\alpha_i^*)}(z_{i,j}) C_k^{(\alpha_i^*)}(z_{i,j}). \\ \Rightarrow \sum_{j=0}^M \omega_j^{(\alpha_i^*)} C_s^{(\alpha_i^*)}(z_{i,j}) f(z_{i,j}) &= \sum_{k=0}^M a_{i,k} \sum_{j=0}^M \omega_j^{(\alpha_i^*)} (\lambda_s^{(\alpha_i^*)} \lambda_k^{(\alpha_i^*)})^{\frac{1}{2}} \phi_s^{(\alpha_i^*)}(z_{i,j}) \phi_k^{(\alpha_i^*)}(z_{i,j}) \\ &= \sum_{k=0}^M a_{i,k} \sum_{j=0}^M \omega_j^{(\alpha_i^*)} (\lambda_s^{(\alpha_i^*)} \lambda_k^{(\alpha_i^*)})^{\frac{1}{2}} \delta_{sk} = a_{i,s} \lambda_s^{(\alpha_i^*)}, \end{aligned}$$

## Chapter 3

---

where  $\{\phi_j^{(\alpha)}(x)\}_{j=0}^M$  are the orthonormal Gegenbauer basis polynomials as defined by Equation (3.A.9). Therefore

$$\begin{aligned} a_{i,s} &= (\lambda_s^{(\alpha_i^*)})^{-1} \sum_{j=0}^M \omega_j^{(\alpha_i^*)} C_s^{(\alpha_i^*)}(z_{i,j}) f(z_{i,j}). \\ \Rightarrow f(x) &\approx \sum_{k=0}^M (\lambda_k^{(\alpha_i^*)})^{-1} \sum_{j=0}^M \omega_j^{(\alpha_i^*)} C_k^{(\alpha_i^*)}(z_{i,j}) C_k^{(\alpha_i^*)}(x) f(z_{i,j}) \\ &= \sum_{k=0}^M \sum_{j=0}^M (\lambda_j^{(\alpha_i^*)})^{-1} \omega_k^{(\alpha_i^*)} C_j^{(\alpha_i^*)}(z_{i,k}) C_j^{(\alpha_i^*)}(x) f(z_{i,k}). \end{aligned} \quad (3.B.12)$$

Hence  $\int_{-1}^{x_i} f(x) dx \approx \sum_{k=0}^M p_{ik}^{(1)}(\alpha_i^*) f(z_{i,k})$ , with  $p_{ik}^{(1)}(\alpha_i^*)$  as defined by Equation (3.20). The quadrature error term (3.23) follows directly from Theorem 3.2.1 on substituting the value of  $\alpha$  with  $\alpha_i^*$ , and expanding the Gegenbauer expansion series up to the  $(M+1)$ th term.

### 3.C Proof of Theorem 3.2.7

Using Lemma 3.2.6 and the error formula (3.23), we can easily derive the following inequality:

$$\begin{aligned} \left| E_M^{(\alpha_i^*)} \right| &\leq \frac{AD_1^{(\alpha_i^*)}}{M^{1/2-\alpha_i^*} (2M/e)^M} \left| \int_{-1}^{x_i} C_{M+1}^{(\alpha_i^*)}(x) dx \right| \\ &\leq \frac{AD_1^{(\alpha_i^*)} (1+x_i)}{M^{1/2-\alpha_i^*} (2M/e)^M} \max_{|x| \leq 1} \left| C_{M+1}^{(\alpha_i^*)}(x) \right| \quad \text{as } M \rightarrow \infty. \end{aligned} \quad (3.C.13)$$

Lemma 3.2.5 yields the asymptotic formula:

$$\max_{|x| \leq 1} \left| C_{M+1}^{(\alpha_i^*)}(x) \right| \approx D_2^{(\alpha_i^*)} (M+1)^{-\alpha_i^*} \quad \forall \alpha_i^* < 0 \text{ as } M \rightarrow \infty. \quad (3.C.14)$$

The proof is established by applying Lemma 3.2.5 for  $\alpha \geq 0$ , and the asymptotic Formula (3.C.14) on the error bound (3.C.13).

### 3.D Proof of Theorem 3.2.8

The proof is divided into two parts: In the first part (I), we show that the Chebyshev quadrature is optimal in the  $L^\infty$ -norm, while in the second part (II), we show that the PMQ converges to the optimal Chebyshev quadrature under the same norm.

## Chapter 3

---

- (I) Denote the bound on the PMQ truncation error for a certain integration node  $x_i$  by  $E_- \forall \alpha_i^* < 0$ , and by  $E_+ \forall \alpha_i^* \geq 0$ . Also let  $\alpha_-; \alpha_+$  be the values of  $\alpha_i^*$  associated with the error bounds  $E_-; E_+$ , respectively. Then Theorem 3.2.7 clearly manifests that the ratio between the PMQ error bounds  $E_-$  and  $E_+$  is equal to a factor of order  $O(M^{-\alpha_+})$ , i.e.

$$\frac{E_-}{E_+} = O(M^{-\alpha_+}) \quad \text{as } M \rightarrow \infty. \quad (3.D.15)$$

Hence the PMQ for negative values of  $\alpha_i^*$  converges faster to the values of the definite integrals than for the positive  $\alpha_i^*$  values. It can be noticed however that for  $\alpha_i^* = 0$ , which corresponds to the Chebyshev quadrature, the error bounds become proportional to each other, so it is essential to determine the relation between the error constants  $B_1^{(\alpha_i^*); B_2^{(\alpha_i^*)}$ . Through Equation (4.7.31) in (Abramowitz and Stegun, 1965) and the Standardization (3.A.2), we can show that

$$\begin{aligned} C_{M+1}^{(\alpha)}(x) &= \sum_{m=0}^{[(M+1)/2]} \frac{(-1)^m (M+1)! \Gamma(2\alpha) \Gamma(M-m+\alpha+1)}{m! (M-2m+1)! \Gamma(\alpha) \Gamma(M+2\alpha+1)} (2x)^{M-2m+1} \\ &= \frac{\Gamma(2\alpha)}{\Gamma(\alpha)} G_M(x) \quad \forall \alpha < 0, \text{ as } M \rightarrow \infty, \end{aligned} \quad (3.D.16)$$

where

$$G_M(x) = \sum_{m=0}^{[(M+1)/2]} \frac{(-1)^m (M+1)! O(\Gamma(M-m+1))}{m! (M-2m+1)! O(\Gamma(M+1))} (2x)^{M-2m+1}. \quad (3.D.17)$$

Therefore  $C_{M+1}^{(\alpha)}(x)$ , for a certain value of  $-1 \leq x \leq 1$ , monotonically decreases for increasing values of  $\alpha$  in the range  $-1/2 < \alpha < 0$ , as  $M \rightarrow \infty$ . Hence the ratio

$$D_2^{(\alpha_i^*)} \approx \max_{|x| \leq 1} \left| C_{M+1}^{(\alpha_i^*)}(x) \right| / (M+1)^{-\alpha_i^*} \quad \forall \alpha_i^* < 0,$$

is also monotonically decreasing for increasing values of  $\alpha_i^*$ . Since  $\lim_{\alpha_i^* \rightarrow 0} D_2^{(\alpha_i^*)} = 1$ , then  $D_2^{(\alpha_i^*)} > 1 \forall \alpha_i^* < 0; B_2^{(\alpha_i^*)} > B_1^{(\alpha_i^*)}$ . Hence

$$\left| E_M^{(0)} \right| < \left| E_M^{(\alpha_i^*)} \right| \quad \forall \alpha_i^* < 0, \text{ as } M \rightarrow \infty. \quad (3.D.18)$$

Relations (3.D.15) & (3.D.18) imply that

$$\left| E_M^{(0)} \right| < \left| E_M^{(\alpha_i^*)} \right| \quad \forall \alpha_i^* \neq 0, \text{ as } M \rightarrow \infty. \quad (3.D.19)$$

## Chapter 3

---

Inequality (3.D.19) is true for any arbitrary set of integration nodes  $\{x_i\}_{i=0}^N$ . Hence we must have

$$0 = \operatorname{argmin}_{\alpha > -1/2} \max_{-1 \leq x \leq 1} \left\| E_M^{(\alpha)}(x, \xi) \right\|, \quad \text{as } M \rightarrow \infty, \quad (3.D.20)$$

and the Chebyshev quadrature is optimal in the  $L^\infty$ -norm, for large values of  $M$ . This completes the proof of the first part.

(II) By Definition (3.16), and Inequality (3.D.19), we have

$$\alpha_i^* = \operatorname{argmin}_{\alpha > -1/2} \eta_{i,M}^2(\alpha) = \operatorname{argmin}_{\alpha > -1/2} |\eta_{i,M}(\alpha)| = \operatorname{argmin}_{\alpha > -1/2} \left| E_M^{(\alpha)}(x_i, \xi_i) \right| = 0 \quad \forall i, \quad (3.D.21)$$

as  $M \rightarrow \infty$ . Hence the PMQ converges to the optimal Chebyshev quadrature for each integration node  $x_i$ , as  $M \rightarrow \infty$ , which completes the proof of the second part.

## 3.E Proof of Lemma 3.2.9

The proof is conveniently divided into three cases: (I)  $j = 0$ , (II)  $j = 1$ ; (III)  $j > 1$ . We shall require the following three well-known properties of the Gamma function:

$$\Gamma(x)\Gamma\left(x + \frac{1}{2}\right) = 2^{1-2x}\sqrt{\pi}\Gamma(2x), \quad (\text{the duplication formula}) \quad (3.E.22a)$$

$$\Gamma(x) \leq x^x e^{1-x} \quad \forall x \geq 1; \quad (3.E.22b)$$

$$\Gamma(x) \geq \left(\frac{x}{e}\right)^{x-1} \quad \forall x \geq 2. \quad (3.E.22c)$$

**Case (I)** For  $j = 0$ : Since

$$\lambda_0^{(\alpha)} = \frac{2^{2\alpha-1}\Gamma^2\left(\alpha + \frac{1}{2}\right)}{\alpha\Gamma(2\alpha)} = \frac{\sqrt{\pi}\Gamma\left(\alpha + \frac{1}{2}\right)}{\alpha\Gamma(\alpha)} = \frac{\sqrt{\pi}\left(\alpha - \frac{1}{2}\right)\Gamma\left(\alpha - \frac{1}{2}\right)}{\alpha\Gamma(\alpha)},$$

using Property (3.E.22a), then

$$\lim_{\alpha \rightarrow \infty} \lambda_0^{(\alpha)} = \sqrt{\pi} \lim_{\alpha \rightarrow \infty} \frac{\Gamma\left(\alpha - \frac{1}{2}\right)}{\Gamma(\alpha)} = 0.$$

**Case (II)** For  $j = 1$ : Here the proof is straightforward since

$$\lambda_1^{(\alpha)} = \frac{2^{2\alpha-1}\Gamma^2\left(\alpha + \frac{1}{2}\right)}{(1+\alpha)\Gamma(1+2\alpha)} = \frac{\alpha}{1+\alpha} \lambda_0^{(\alpha)} \rightarrow 0 \quad \text{as } \alpha \rightarrow \infty.$$

## Chapter 3

---

**Case (III)** For  $j > 1$ : Since

$$\Gamma\left(\alpha + \frac{1}{2}\right) \leq \left(\alpha + \frac{1}{2}\right)^{\alpha + \frac{1}{2}} e^{\frac{1}{2} - \alpha} \quad \forall \alpha \geq \frac{1}{2}, \quad (3.E.23)$$

$$\Gamma(j + 2\alpha) \geq \left(\frac{j + 2\alpha}{e}\right)^{j + 2\alpha - 1} \quad \forall \alpha \geq 1 - \frac{j}{2}, \quad (3.E.24)$$

then

$$\frac{\Gamma^2\left(\alpha + \frac{1}{2}\right)}{\Gamma(j + 2\alpha)} \leq \frac{e^j (2\alpha + 1)^{2\alpha + 1}}{2^{2\alpha + 1} (j + 2\alpha)^{j + 2\alpha - 1}} \quad \forall \alpha \geq \frac{1}{2}.$$

Hence

$$\lambda_j^{(\alpha)} \leq \frac{j! e^j (2\alpha + 1)}{4(j + \alpha)(j + 2\alpha)^{j-1}} \left(\frac{2\alpha + 1}{2\alpha + j}\right)^{2\alpha} \rightarrow 0 \quad \text{as } \alpha \rightarrow \infty.$$

### 3.F Proof of Theorem 3.2.10

Assume that  $M$  is a fixed number,  $\alpha_i^* \rightarrow \infty$ ;  $|f(x)| \leq A \in \mathbb{R}^+$ , for some smooth integrand function  $f(x)$ . Since  $(\lambda_k^{(\alpha_i^*)})^{-1} \gg 1$  as proven by Lemma 3.2.9;  $C_k^{(\alpha_i^*)}(z_{i,j}) < 1$  for all  $i, j, k$ , then  $(\lambda_k^{(\alpha_i^*)})^{-1} (C_k^{(\alpha_i^*)}(z_{i,j}))^2 = O\left((\lambda_k^{(\alpha_i^*)})^{-1}\right)$ . From Equations (3.21); (3.B.12), we have

$$\begin{aligned} \mathcal{K}_{i,s} &= \left| \sum_{j=0}^M \omega_j^{(\alpha_i^*)} C_s^{(\alpha_i^*)}(z_{ij}) f(z_{ij}) \right| < A \sum_{j=0}^M \left| \omega_j^{(\alpha_i^*)} \right| \\ \Rightarrow \text{Sup}_s \mathcal{K}_{i,s} &= A \sum_{j=0}^M \left( \frac{1}{\sum_{k=0}^M O\left((\lambda_k^{(\alpha_i^*)})^{-1}\right)} \right) = \frac{(M+1)A}{\sum_{k=0}^M O\left((\lambda_k^{(\alpha_i^*)})^{-1}\right)}, \end{aligned}$$

which decays exponentially fast. Since  $\text{Sup}_s \mathcal{K}_{i,s}$  converges rapidly to zero faster than the growth rate of  $(\lambda_s^{(\alpha_i^*)})^{-1}$ , for increasing values of  $\alpha_i^*$ , the magnitudes of the Gegenbauer coefficients  $a_{i,s}$  vanish very quickly as  $\alpha_i^* \rightarrow \infty$ . This completes the proof of the theory.

## Chapter 3

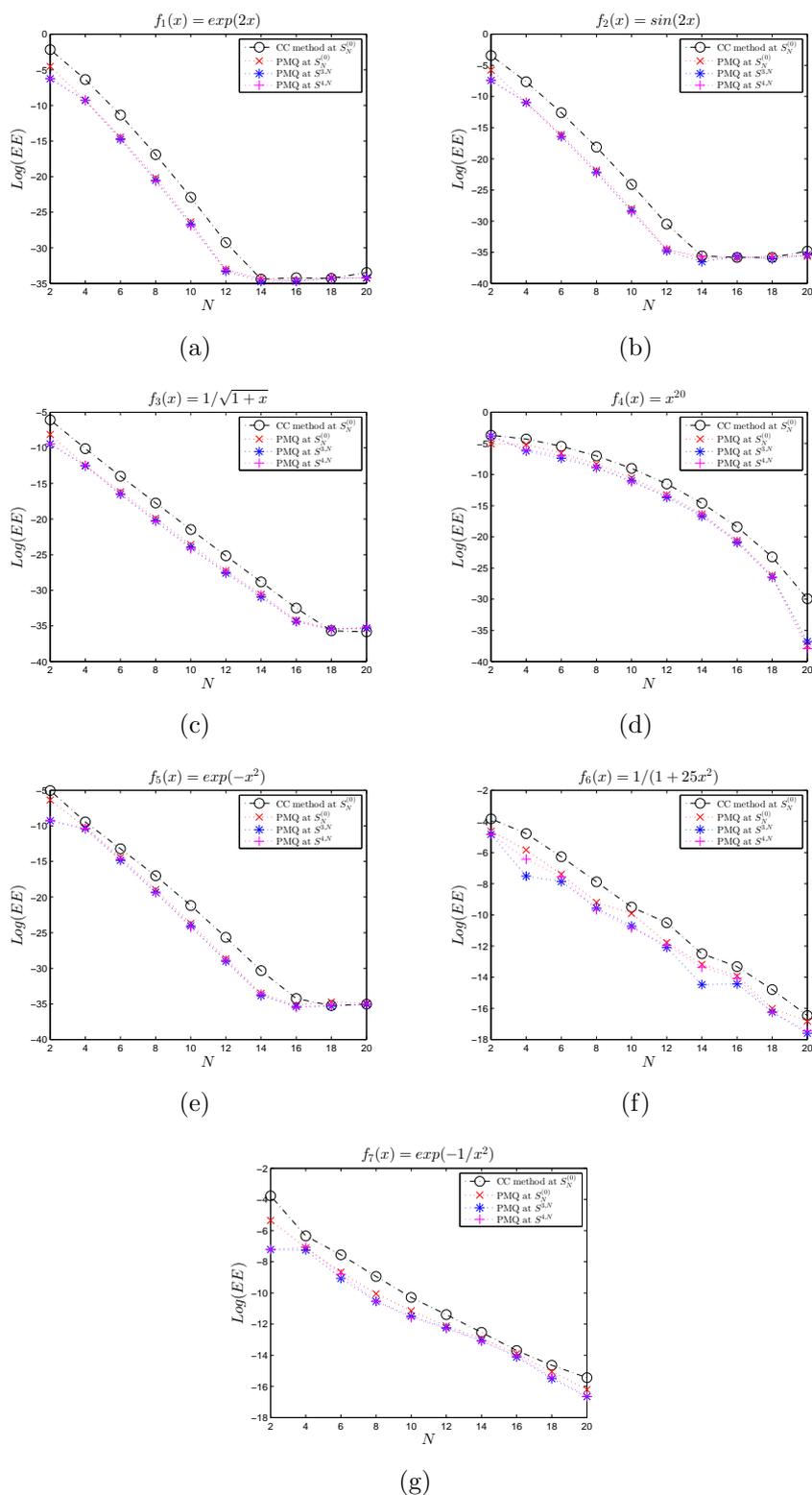


Figure 4:  $\text{Log}(EE)$  in floating-point arithmetic for the PMQ and the CC quadrature versus  $N = 2, 4, \dots, 20$ , for the seven test functions  $\{f_i\}_{i=1}^7$  on  $[0, 1]$ . The results of the CC method are reported at  $S_N^{(0)}$ . The results of the PMQ are reported at  $S_N^{(0)}, S^{3,N}; S^{4,N}$ .

## Chapter 3

---

### 3.G Algorithm 2.1

---

**Algorithm 2.1** Construction of the P-matrix for any general set of integration points

---

Input Integer numbers  $N, M, M_{\max}$ ; positive real number  $r$ ; the set of the integration nodes  $\{x_i\}_{i=0}^N$ ; relatively small positive number  $\varepsilon$ .

Output The P-matrix  $P = (p_{ij}), i = 0, \dots, N; j = 0, \dots, M$ .

Step 1 If  $M > M_{\max}$  then

If  $M = N$  then set  $P = \hat{P}$  with  $\alpha = 0.5$ ;  
Output  $P$ ; Stop.

Else calculate the Legendre-Gauss points  $\hat{x}_j \in S_{M+1}^{(0.5)}$ ;

Calculate  $L_j(\hat{x}_m); \lambda_j^{(0.5)}; \omega_j^{(0.5)} \forall 0 \leq j, m \leq M$ ;

Calculate  $\int_{-1}^{x_i} L_j(x) dx \forall j = 0, \dots, M$ ;

Calculate  $p_{ij} = \sum_{k=0}^M (\lambda_k^{(0.5)})^{-1} \omega_j^{(0.5)} L_k(\hat{x}_j) \int_{-1}^{x_i} L_k(x) dx \forall i = 0, \dots, N; j = 0, \dots, M$ ;

Output  $P$ ; Stop.

Step 2 Set the counter  $i = 0$ .

Step 3 While  $i \leq N$  do Steps 4-7:

Step 4 Solve  $\alpha_i^* = \operatorname{argmin}_{-1/2 < \alpha \leq r} \eta_{i,M}^2(\alpha)$ .

Step 5 If  $\alpha_i^* \in (-0.5, -0.5 + \varepsilon)$  choose  $\alpha_i^* \in \{-0.5 + \varepsilon, 0.5\}$ .

Step 6 Calculate the zeros points  $z_{i,j}, j = 0, \dots, M$ , of the Gegenbauer polynomial  $C_{M+1}^{(\alpha_i^*)}(x)$ ;

Calculate  $C_j^{(\alpha_i^*)}(z_{i,m}); \lambda_j^{(\alpha_i^*)}; \omega_j^{(\alpha_i^*)} \forall 0 \leq j, m \leq M$ ;

Calculate  $\int_{-1}^{x_i} C_j^{(\alpha_i^*)}(x) dx \forall j = 0, \dots, M$ ;

Calculate  $p_{ij} \forall 0 \leq j \leq M$  as defined by Equation (3.20).

Step 7 Set  $i = i + 1$ ; go to Step 3.

Step 8 Output  $P$ ; Stop.

---

## Chapter 3

---

### 3.H Algorithm 2.2

---

**Algorithm 2.2** Reduced construction of the P-matrix for any symmetric set of integration points

---

Input Integer numbers  $N, M, M_{\max}$  where  $N$  is even; positive real number  $r$ ; the set of the integration nodes  $\{x_i\}_{i=0}^N$ ; relatively small positive number  $\varepsilon$ .

Output The P-matrix  $P = (p_{ij}), i = 0, \dots, N; j = 0, \dots, M$ .

Step 1 Apply Step 1 in Algorithm 2.1.

Step 2 Set the counter  $i = 0$ .

Step 3 While  $i \leq N$  do Steps 4-9:

Step 4 If  $i \leq N/2$  do Steps 5-7:

Step 5 Solve  $\alpha_i^* = \underset{-1/2 < \alpha \leq r}{\operatorname{argmin}} \eta_{i,M}^2(\alpha)$ .

Step 6 If  $\alpha_i^* \in (-0.5, -0.5 + \varepsilon)$  choose  $\alpha_i^* \in \{-0.5 + \varepsilon, 0.5\}$ .

Step 7 Calculate the zeros points  $z_{i,j}, j = 0, \dots, M$ , of the Gegenbauer polynomial  $C_{M+1}^{(\alpha_i^*)}(x)$ ;

Calculate  $\beta_{1,i} = C_j^{(\alpha_i^*)}(z_{i,m}); \beta_{2,i} = \lambda_j^{(\alpha_i^*)}; \beta_{3,i} = \omega_j^{(\alpha_i^*)} \forall 0 \leq j, m \leq M$ ;

Set  $\alpha_{N-i}^* = \alpha_i^*; z_{N-i,j} = z_{i,j}; \beta_{1,N-i} = \beta_{1,i}; \beta_{2,N-i} = \beta_{2,i}; \beta_{3,N-i} = \beta_{3,i} \forall j = 0, \dots, M$ .

Step 8 Calculate  $\int_{-1}^{x_i} C_j^{(\alpha_i^*)}(x) dx \forall j = 0, \dots, M$ ;

Calculate  $p_{ij} \forall 0 \leq j \leq M$  as defined by Equation (3.20).

Step 9 Set  $i = i + 1$ ; go to Step 3.

Step 10 Output  $P$ ; Stop.

---

## PART B: Suggested Declaration for Thesis Chapter

Monash University

### Declaration for Thesis Chapter 4

#### Declaration by candidate

In the case of Chapter 4, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
The author of the key ideas, organization, development, and writing up of the article	95

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided valuable comments and aided proofreading	

Candidate's  
Signature

	Date 11/05/2013
---	--------------------

#### Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication; and writing up of the article 95
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)

Signature 1

	Date 14/5/13
---	--------------

.....

This page is intentionally left blank

## *Chapter 4*

# *On the Optimization of Gegenbauer Operational Matrix of Integration*

**Chapter 4 is based on the published article Elgindy, K. T., Smith-Miles, K. A., 1 December 2012. On the optimization of Gegenbauer operational matrix of integration. *Advances in Computational Mathematics*, Springer US, 1–14. DOI 10.1007/s10444-012-9289-5.**

**Abstract.** *The theory of Gegenbauer (ultraspherical) polynomial approximation has received considerable attention in recent decades. In particular, the Gegenbauer polynomials have been applied extensively in the resolution of the Gibbs phenomenon, construction of numerical quadratures, solution of ordinary and partial differential equations, integral and integro-differential equations, optimal control problems, etc. To achieve better solution approximations, some methods presented in the literature apply the Gegenbauer operational matrix of integration for approximating the integral operations, and recast many of the aforementioned problems into unconstrained/constrained optimization problems. The Gegenbauer parameter  $\alpha$  associated with the Gegenbauer polynomials is then added as an extra unknown variable to be optimized in the resulting optimization problem as an attempt to optimize its value rather than choosing a random value. This issue is addressed in this chapter as we prove theoretically that it is invalid. In particular, we provide a solid mathematical proof demonstrating that optimizing the Gegenbauer operational matrix of integration for the solution of various mathematical problems by recasting them into equivalent optimization problems with  $\alpha$  added as an extra optimization variable violates the discrete Gegenbauer orthonormality relation, and may in turn produce false solution approximations.*

**Keyword.** *Gegenbauer integration matrix, Gegenbauer operational matrix of integration, Gegenbauer polynomials, Spectral methods.*

References are considered at the end of the thesis.

# Chapter 4

## On the Optimization of Gegenbauer Operational Matrix of Integration

### 4.1 Introduction

ODEs, integral equations, integro-differential equations, and optimal control problems are encountered in the modeling of many real life problems. Most of these problems are usually nonlinear, and finding analytical solutions is almost impossible in many cases. The elegant class of methods known as the spectral methods can provide excellent approximations for problems exhibiting smooth solutions; cf. (Boyd, 2000; Canuto et al., 1988, 2006; Elbarbary, 2007; Elgindy, 2009; Fornberg, 1996; Gottlieb and Orszag, 1977; Guo, 1998; Hesthaven et al., 2007; Mercier, 1989; Tian, 1989; Trefethen, 2000). Moreover, this class of spatial discretizations is more favorable than the finite element and the finite difference methods for several reasons such as their promise of exponential convergence yielding an error of order  $O(1/N^N)$  for expansions up to the  $N^{th}$  term (Gottlieb and Orszag, 1977); they require less memory storage than alternative methods (Boyd, 2000), etc.

In a typical spectral method for solving ODEs, the unknown solutions are approximated by truncated spectral expansion series, and the derivatives of the unknown solutions are approximated by the spectral differentiation matrices. Although linear ODEs are ultimately transformed into linear systems of algebraic equations, which can be solved by efficient linear system solvers, spectral differentiation matrices are known to be severely ill-conditioned (Funaro, 1987), and are prone to large round-off errors. Consequently, these effects cause degradation of the observed precision (Greengard, 1991; Tang and Trummer, 1996), and render the development of efficient preconditioners crucial and a necessity in many

## Chapter 4

---

cases, as the procedure involves the solution of very ill-conditioned linear system of equations (Elbarbary, 2006; Hesthaven, 1998). Alternatively, other spectral methods presented in the literature transform the linear mathematical problems into algebraic linear systems of equations through *spectral integration matrices*; cf. (El-Gendi, 1969; Elbarbary, 2006, 2007; Elgindy, 2009; Elgindy and Smith-Miles, 2013c; Ghoreishi and Hosseini, 2008; Mai-Duy and Tanner, 2007; Mihaila and Mihaila, 2002; Tian, 1989). The resulting algebraic equations systems can then be solved using efficient and well-known linear system solvers. This approach overcomes the drawbacks of applying the spectral differentiation matrices as spectral integration matrices are known to be well-conditioned operators (Greengard, 1991; Lundbladh et al., 1992); their well-conditioning is essentially unaffected for increasing number of grid points (Elgindy, 2009; Elgindy and Smith-Miles, 2013c).

Since the Gegenbauer polynomials form a complete basis system in  $L^2([-1, 1], (1-x^2)^{\alpha-1/2})$ , in a classical Gegenbauer spectral method for solving an ODE, the unknown solution  $y(x)$  is expanded in a finite series of the smooth basis polynomials  $C_j^{(\alpha)}(x)$  in the form

$$y(x) \approx \sum_{j=0}^N a_j C_j^{(\alpha)}(x), \quad (4.1)$$

where  $C_j^{(\alpha)}(x)$  is the Gegenbauer polynomial of degree  $j$  and associated with the parameter  $\alpha$ ; cf. Section 4.2;  $\{a_j\}_{j=0}^N$  are the Gegenbauer coefficients of the solution function  $y(x)$ . The solution/collocation nodes set is frequently chosen to be the set of Gegenbauer-Gauss nodes (the zeros of the Gegenbauer polynomial  $C_{N+1}^{(\alpha)}(x)$ ) defined by:

$$S_N^{(\alpha)} = \{x_k | C_{N+1}^{(\alpha)}(x_k) = 0, k = 0, \dots, N\},$$

for reasons that can be probed from the area of approximation theory (Trefethen, 2000). The integral operations can be approximated using the Gegenbauer operational matrices of integration (Gegenbauer integration matrices); cf. Section 4.2, after recasting the ODE into its integral formulation. Since the construction of the Gegenbauer integration matrix, as we shall describe later, depends primarily on the choice of the Gegenbauer parameter  $\alpha$ , an intuitive idea is to optimize the value of  $\alpha$  rather than opting for a random choice in the interval  $(-1/2, \infty)$ . This raises the intriguing question of “*which value of the Gegenbauer parameter  $\alpha$  is optimal for a Gegenbauer integration matrix to best approximate the solution of the problem?*” There have been some attempts in the literature to optimize the Gegenbauer integration matrix by transforming various mathematical problems into unconstrained/constrained optimization problems, which are

## Chapter 4

---

then solved using some of the standard optimization methods. This technique has been applied by El-Hawary et al. (2003) in the solution of optimal control problems by transforming the latter into constrained nonlinear programming problems. Later El-Kady et al. (2009) applied a similar method for the solution of linear integro-differential equations by transforming them into unconstrained optimization problems. The reason behind these transformations was essentially to add the Gegenbauer parameter  $\alpha$  as an extra variable to be optimized to seek better approximations rather than applying the Gegenbauer integration matrices with a preselected  $\alpha$  value. However our theoretical analysis presented in Section 4.4 shows a deficiency arising during the optimization procedure. In fact, it can be shown that it is mathematically erroneous to optimize the Gegenbauer parameter  $\alpha$  after transforming the original mathematical problem into an unconstrained/constrained optimization problem, as it violates the discrete Gegenbauer orthonormality relation; hence may produce false solution approximations. We highlight how this tempting approach to optimize the Gegenbauer integration matrix is deceptive. This chapter is important as it rebuts the aforementioned solution methods. Moreover, the chapter represents a barrier against any attempts in the future to optimize the Jacobi integration matrix or other class of functions integration matrices based on similar ideas. The remaining part of the chapter is organized as follows: In the following section we shall briefly discuss some properties of the Gegenbauer polynomials and their associated integration matrices. In Section 4.3 we shall describe the procedure of adding the Gegenbauer parameter  $\alpha$  as an extra variable to be optimized through two simple examples. A solid mathematical proof which outlines the pitfalls of optimizing the Gegenbauer integration matrix by transforming the mathematical problem into an equivalent optimization problem is presented in Section 4.4. Finally, some concluding remarks on the optimization of the Gegenbauer integration matrix and a viable alternative method are presented in Section 4.5.

### 4.2 Preliminary Definitions and Properties

The Gegenbauer polynomials associated with the parameter  $\alpha > -1/2$  appear as eigensolutions to the following singular Sturm-Liouville problem in the finite domain  $[-1, 1]$  (Szegő, 1975):

$$\frac{d}{dx}(1-x^2)^{\alpha+\frac{1}{2}} \frac{dC_n^{(\alpha)}(x)}{dx} + n(n+2\alpha)(1-x^2)^{\alpha-\frac{1}{2}} C_n^{(\alpha)}(x) = 0,$$

with the first two being

$$C_0^{(\alpha)}(x) = 1, \quad C_1^{(\alpha)}(x) = 2\alpha x,$$

## Chapter 4

---

while the remaining polynomials are given through the three-term recursion formula

$$(n+1)C_{n+1}^{(\alpha)}(x) = 2(n+\alpha)x C_n^{(\alpha)}(x) - (n+2\alpha-1)C_{n-1}^{(\alpha)}(x), \quad n = 1, 2, \dots$$

The zeros of the Gegenbauer polynomials are obtained numerically, and their behavior has been of interest because of their nice electrostatic interpretation (Hendriksen and van Rossum, 1988), and of their important role as nodes of Gaussian quadrature formulae (Area et al., 2004). The weight function for the Gegenbauer polynomials is the even function  $(1-x^2)^{\alpha-1/2}$  (Abramowitz and Stegun, 1965); the orthogonality relation of the Gegenbauer polynomials is given by

$$\int_{-1}^1 (1-x^2)^{\alpha-1/2} C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = h_n^{(\alpha)} \delta_{mn}, \quad (4.2)$$

where

$$h_n^{(\alpha)} = \frac{2^{1-2\alpha} \pi \Gamma(n+2\alpha)}{n!(n+\alpha)\Gamma^2(\alpha)}; \quad (4.3)$$

$\delta_{mn}$  is the Kronecker delta function. It is important to note that the form of the Gegenbauer polynomials is in fact not unique, and depends on how they are standardized. Doha (1990) standardized the Gegenbauer polynomials so that

$$C_n^{(\alpha)}(1) = 1, \quad n = 0, 1, 2, \dots \quad (4.4)$$

This standardization establishes the useful relations that  $C_n^{(0)}(x)$  becomes identical with the Chebyshev polynomial of the first kind  $T_n(x)$ ,  $C_n^{(1/2)}(x)$  is the Legendre polynomial  $L_n(x)$ ;  $C_n^{(1)}(x)$  is equal to  $(1/(n+1))U_n(x)$ , where  $U_n(x)$  is the Chebyshev polynomial of the second type. Consequently, the Gegenbauer polynomials can be generated by Rodrigues' formula in the following form:

$$C_n^{(\alpha)}(x) = \left(-\frac{1}{2}\right)^n \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} (1-x^2)^{\frac{1}{2}-\alpha} \frac{d^n}{dx^n} (1-x^2)^{n+\alpha-\frac{1}{2}}; \quad (4.5)$$

they satisfy the orthogonality relation (Elgindy and Smith-Miles, 2013b)

$$\int_{-1}^1 (1-x^2)^{\alpha-1/2} C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = \lambda_n^{(\alpha)} \delta_{mn}, \quad (4.6)$$

where

$$\lambda_n^{(\alpha)} = \frac{2^{2\alpha-1} n! \Gamma^2(\alpha + \frac{1}{2})}{(n+\alpha)\Gamma(n+2\alpha)}. \quad (4.7)$$

For the rest of the chapter, when we refer to the Gegenbauer polynomials, we shall mean the Gegenbauer polynomials standardized so that equation (4.4) is

## Chapter 4

---

satisfied. The definite integration of a function  $f \in C[-1, 1]$  is approximated by performing integration on the finite Gegenbauer expansion series, and the resulting integration approximations for the Gegenbauer-Gauss integration nodes  $x_i \in S_N^{(\alpha)}$  can be expressed in a matrix-vector multiplication form  $I = QF$ , where

$$I = \left[ \int_{-1}^{x_0} f(x) dx, \int_{-1}^{x_1} f(x) dx, \dots, \int_{-1}^{x_N} f(x) dx \right]^T, \quad Q = (q_{ij}), \quad 0 \leq i, j \leq N;$$

$$F = [f(x_0), f(x_1), \dots, f(x_N)]^T.$$

The operational matrix  $Q$  is referred to as the Gegenbauer integration matrix. One approach for constructing the  $Q$ -matrix can be described by the following theorem (El-Hawary et al., 2000):

**Theorem 4.2.1.** *Let  $f(x)$  be approximated by the Gegenbauer polynomials;  $x_k \in S_N^{(\alpha)}$ , then there exist a matrix  $Q = (q_{ij})$ ,  $i, j = 0, \dots, N$ ; and a number  $\xi = \xi(x) \in [-1, 1]$  satisfying*

$$\int_{-1}^{x_i} f(x) dx = \sum_{k=0}^N q_{ik}(\alpha) f(x_k) + E_N^{(\alpha)}(x_i, \xi), \quad (4.8)$$

where

$$q_{ik}(\alpha) = \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k) \int_{-1}^{x_i} C_j^{(\alpha)}(x) dx, \quad (4.9)$$

$$(\omega_k^{(\alpha)})^{-1} = \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} (C_j^{(\alpha)}(x_k))^2, \quad (4.10a)$$

$$\lambda_j^{(\alpha)} = 2^{j+2\alpha+\tau} j! \frac{\Gamma(\alpha + \frac{1}{2}) \Gamma(j + \alpha + \frac{1}{2})}{\Gamma(2j + 2\alpha + 1)} K_j^{(\alpha)}, \quad (4.10b)$$

$$\tau = \begin{cases} 1, & \text{if } \alpha = j = 0, \\ 0, & \text{otherwise,} \end{cases} \quad (4.10c)$$

$$K_j^{(\alpha)} = 2^j \frac{\Gamma(j + \alpha) \Gamma(2\alpha + 1)}{\Gamma(j + 2\alpha) \Gamma(\alpha + 1)}; \quad (4.10d)$$

$$E_N^{(\alpha)}(x, \xi) = \frac{f^{(N+1)}(\xi)}{(N+1)! K_{N+1}^{(\alpha)}} \int_{-1}^x C_{N+1}^{(\alpha)}(x) dx. \quad (4.11)$$

*Proof.* See (El-Hawary et al., 2000). □

## Chapter 4

---

Equation (4.9) defines the elements of the Q-matrix calculated at  $S_N^{(\alpha)}$ , Equations (4.10a)-(4.10d) define the required parameters; Equation (4.11) defines the error term. For further properties of the Gegenbauer polynomials, the reader may consult (Abramowitz and Stegun, 1965; Andrews et al., 1999; Bayin, 2006; Szegő, 1975), and the literature quoted there. In the following section, we shall discuss the issue of achieving higher-order approximations to the solutions of diverse mathematical problems by optimizing the Gegenbauer parameter  $\alpha$ ; thus optimizing the Q-matrix.

### 4.3 Solving Various Mathematical Problems by Optimizing the Gegenbauer Integration Matrix

Suppose that  $y(x) \in C^\infty[-1, 1]$  is the unknown solution of a certain mathematical problem, generally an ODE, integro-differential equation, or an integral equation. We seek the values of this solution at the Gegenbauer-Gauss nodes  $x_i \in S_N^{(\alpha)}$ . ODEs and integro-differential equations can be recast into integral equations by direct integration, or by substituting the highest order derivative, say  $y^{(N)}(x)$ , with some unknown function, say  $\phi(x)$ , and solving the problem for the unknown function  $\phi(x)$ . The unknown solution  $y(x)$  and its derivatives up to the  $(N-1)^{th}$ -order derivative can then be obtained by successive integrations of the function  $\phi(x)$ . We shall discuss the issue of achieving higher-order approximations by optimizing the Gegenbauer integration matrix for integral equations only. For simplicity, and without loss of generality, consider the following integral equation:

$$y(x) + \int_{-1}^x f(x)y(x)dx = g(x), \quad (4.12)$$

for some continuous functions  $f(x)$  and  $g(x)$  on  $[-1, 1]$ . Expanding the solution  $y(x)$  by the Gegenbauer expansion series (4.1), and substituting into the integral equation (4.12) transform the latter into a system of linear algebraic equations of the following form:

$$I_i(w) = w_i - g_i + \sum_{j=0}^N q_{ij}^{(1)} f_j w_j = 0, \quad i = 0, \dots, N, \quad (4.13)$$

where  $w = [w_0, w_1, \dots, w_N]^T$ ,  $w_i \approx y(x_i)$ ,  $g_i = g(x_i)$ ,  $f_i = f(x_i)$ ,  $q_{ij}^{(1)}$ ;  $0 \leq i, j \leq N$  are the elements of the first-order Q-matrix. Since the form of the Q-matrix depends on the value of the Gegenbauer parameter  $\alpha$ , the latter may be added

## Chapter 4

---

as an extra unknown; consequently, Equations (4.13) may be reformulated as follows:

$$I_i(w, \alpha) = w_i - g_i + \sum_{j=0}^N q_{ij}^{(1)}(\alpha) f_j w_j = 0, \quad i = 0, \dots, N. \quad (4.14)$$

The solution of the linear system (4.14) is equivalent to the solution of the following constrained optimization problem:

$$\min_{w, \alpha} J(w, \alpha) = \sum_{i=0}^N I_i^2(w, \alpha), \quad (4.15a)$$

$$\text{subject to } \alpha > -1/2, \quad (4.15b)$$

where the objective function  $J$  is a multivariate function of  $(N+2)$  unknowns. To satisfy the constraint  $\alpha > -1/2$ , the following change of variable can be employed:

$$\alpha = e^{(t^2 + \varepsilon)} - \frac{3}{2}, \quad 0 < \varepsilon \ll 1, \quad (4.16)$$

transforming the constrained optimization problem (4.15) into an unconstrained optimization problem in the unknown parameter  $t$ . Considering the unknown solution vector  $w$  and  $\alpha$  as the free variables, one may think that he can apply any standard optimization method to determine the optimal  $\alpha^*$  value such that the corresponding solution approximation  $w^*$  is the most accurate approximation to the unknown solution  $y(x)$  of the original problem (4.12). However, the literature is lacking examples where this has been conveniently done. Instead, we find studies where the authors have acknowledged that the accuracy of the results is quite sensitive, and depends on the right choice of the initial guess of the Gegenbauer parameter  $\alpha$ . For instance, the authors in (El-Kady et al., 2009) undertook lots of trial and error studies, and started with arbitrary choices of  $\alpha = 0.503, 0.345, -0.311; -0.322$  for solving some linear integro-differential equations test problems. When we analyze the resulting optimization problem, we can find a clear explanation for the sensitivity in the initial guesses of the Gegenbauer parameter  $\alpha$ . In particular, it can be shown that there is a coupling of the solution nodes  $\{x_k\}_{k=0}^N$  and the value of  $\alpha$ , which means that we are not completely free to choose any arbitrary  $\alpha$  independently in the search space. We need to maintain that coupling in order to obtain accurate approximations for the Gegenbauer coefficients  $\{a_j\}_{j=0}^N$  in the Gegenbauer expansion series (4.1); consequently obtain correct formulation for the Q-matrix entries (4.9). The previous studies found that the accuracy of the results is so sensitive to the initial value of  $\alpha$ , because any attempt to search for the optimal  $\alpha^*$  value may break that coupling, and “violates the discrete Gegenbauer orthonormality relation.” As a result, the obtained optimization problem is in fact a false optimization problem

## Chapter 4

---

that is a poor approximation to the original problem. Hence, although on the surface it may sound reasonable to optimize the Gegenbauer integration matrix by optimizing the Gegenbauer parameter  $\alpha$  through an equivalent unconstrained optimization problem, the addition of the Gegenbauer parameter  $\alpha$  as an extra variable to be optimized is indisputably incorrect.

In the above argument, we described the addition of the Gegenbauer parameter  $\alpha$  as an extra unknown variable to be optimized to best approximate the solution of an integral equation problem. In general, this method may be misused in many applications as there are many mathematical problems which can be recast as nonlinear optimization problems. In optimal control theory, for instance, it is known that a class of methods called the direct methods converts the continuous optimal control problem into a finite dimensional constrained optimization problem. Approximating the state and the control variables by the Gegenbauer expansion series transforms the optimal control problem into a constrained optimization problem of the form:

$$\text{minimize } J(a, b) \tag{4.17a}$$

$$\text{subject to } h_i(a, b) = 0, \tag{4.17b}$$

$$g_j(a, b) \leq 0, \tag{4.17c}$$

for some numbers  $i, j \in \mathbb{Z}^+$ , where  $J$  is the cost function,  $h_i; g_j$  are some equality and inequality constraints functions, and  $a; b$  are the Gegenbauer coefficients of the state and the control variables. If we add the Gegenbauer parameter  $\alpha$  as an extra unknown variable to be optimized, the constrained optimization problem (4.17) may then be reformulated in the following form:

$$\text{minimize } J(a, b, \alpha) \tag{4.18a}$$

$$\text{subject to } h_i(a, b, \alpha) = 0, \tag{4.18b}$$

$$g_j(a, b, \alpha) \leq 0, \tag{4.18c}$$

$$\alpha > -1/2, \tag{4.18d}$$

cf. (El-Hawary et al., 2003), where the change of variable (4.16) is again applied to satisfy constraint (4.18d). To conclude, we can demonstrate theoretically that these presented methods are not possible, and this gives a clarification for the extreme sensitivity in the initial guesses of the parameter  $\alpha$ . We shall present the mathematical proof in the following section.

### 4.4 The Mathematical Proof

Suppose that the unknown solution  $y(x)$  is approximated by the Gegenbauer expansion series (4.1);  $x_k^{(\alpha)} \in S_N^{(\alpha)} \forall k$ . Following the approach presented by El-Hawary et al. (2000), the discrete representation of the Gegenbauer coefficients

## Chapter 4

---

$\{a_j\}_{j=0}^N$  can be obtained via a discrete least squares fitting at the elements of  $S_N^{(\alpha)}$  in the form:

$$a_j = (\lambda_j^{(\alpha)})^{-1} \sum_{k=0}^N \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k^{(\alpha)}) y(x_k^{(\alpha)}), \quad j = 0, \dots, N. \quad (4.19)$$

Hence Equation (4.1) can be reformulated as follows:

$$y(x) \approx \sum_{j=0}^N \sum_{k=0}^N (\lambda_j^{(\alpha)})^{-1} \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k^{(\alpha)}) C_j^{(\alpha)}(x) y(x_k^{(\alpha)}). \quad (4.20)$$

Therefore

$$\begin{aligned} \int_{-1}^{x_i^{(\alpha)}} y(x) dx &\approx \sum_{k=0}^N \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k^{(\alpha)}) y(x_k^{(\alpha)}) \int_{-1}^{x_i^{(\alpha)}} C_j^{(\alpha)}(x) dx \\ &= \sum_{k=0}^N q_{ik}^{(1)}(\alpha) y(x_k^{(\alpha)}), \end{aligned} \quad (4.21)$$

where

$$q_{ik}^{(1)}(\alpha) = \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k^{(\alpha)}) \int_{-1}^{x_i^{(\alpha)}} C_j^{(\alpha)}(x) dx, \quad 0 \leq i, k \leq N, \quad (4.22)$$

are the elements of the Q-matrix. Now suppose that  $\alpha$  is added as an extra unknown variable in the optimization problem as in Problems (4.15) and (4.18), and say that  $\alpha_0$  is the initial guess of the optimal  $\alpha^*$ , and the solution points  $x_k^{(\alpha)} \in S_N^{(\alpha)}$  for each  $k$ . In this case Equation (4.19) can be written as

$$a_j = (\lambda_j^{(\alpha_0)})^{-1} \sum_{k=0}^N \omega_k^{(\alpha_0)} C_j^{(\alpha_0)}(x_k^{(\alpha_0)}) y(x_k^{(\alpha_0)}), \quad j = 0, \dots, N; \quad (4.23)$$

Equations (4.21) and (4.22) become

$$\int_{-1}^{x_i^{(\alpha_0)}} y(x) dx \approx \sum_{k=0}^N q_{ik}^{(1)}(\alpha_0) y(x_k^{(\alpha_0)}); \quad (4.24)$$

$$q_{ik}^{(1)}(\alpha_0) = \sum_{j=0}^N (\lambda_j^{(\alpha_0)})^{-1} \omega_k^{(\alpha_0)} C_j^{(\alpha_0)}(x_k^{(\alpha_0)}) \int_{-1}^{x_i^{(\alpha_0)}} C_j^{(\alpha_0)}(x) dx, \quad 0 \leq i, k \leq N. \quad (4.25)$$

## Chapter 4

---

The value of  $\alpha = \alpha_0$  is then modified in every iteration until it converges to the optimal value  $\alpha^*$ . Suppose that in iteration  $m$ , the value of  $\alpha_0$  changed to  $\alpha_m$  such that  $\alpha_m \neq \alpha_0$ , then Equations (4.24) and (4.25) become

$$\int_{-1}^{x_i^{(\alpha_0)}} y(x) dx \approx \sum_{k=0}^N q_{ik}^{(1)}(\alpha_m) y(x_k^{(\alpha_0)}); \quad (4.26)$$

$$q_{ik}^{(1)}(\alpha_m) = \sum_{j=0}^N (\lambda_j^{(\alpha_m)})^{-1} \omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)}) \int_{-1}^{x_i^{(\alpha_0)}} C_j^{(\alpha_m)}(x) dx, \quad 0 \leq i, k \leq N. \quad (4.27)$$

Let  $x_i^{(\alpha_0)} = x \in [-1, 1]$ , then Equation (4.26) can be written as

$$\int_{-1}^x y(x) dx \approx \sum_{k=0}^N \sum_{j=0}^N (\lambda_j^{(\alpha_m)})^{-1} \omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)}) y(x_k^{(\alpha_0)}) \int_{-1}^x C_j^{(\alpha_m)}(x) dx. \quad (4.28)$$

Differentiating both sides with respect to  $x$  yields

$$y(x) \approx \sum_{k=0}^N \sum_{j=0}^N (\lambda_j^{(\alpha_m)})^{-1} \omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)}) y(x_k^{(\alpha_0)}) C_j^{(\alpha_m)}(x). \quad (4.29)$$

$$\Rightarrow a_j = (\lambda_j^{(\alpha_m)})^{-1} \sum_{k=0}^N \omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)}) y(x_k^{(\alpha_0)}), \quad j = 0, \dots, N. \quad (4.30)$$

Hence Equation (4.30) gives the Gegenbauer coefficients  $a_j$  at iteration  $m$ , which produce the Gegenbauer integration entries (4.27). Our concern now is to prove that Equation (4.30) is false; consequently, Equation (4.27) is erroneous. To prove our claim, let us introduce the orthonormal Gegenbauer polynomial  $\phi_j^{(\alpha)}(x)$  of degree  $j$  and associated with the Gegenbauer parameter  $\alpha$  by the following relation:

$$\phi_j^{(\alpha)}(x) = (\lambda_j^{(\alpha)})^{-\frac{1}{2}} C_j^{(\alpha)}(x). \quad (4.31)$$

The orthonormal Gegenbauer polynomials  $\phi_j^{(\alpha)}(x)$  satisfy the following discrete orthonormality relation:

$$\sum_{k=0}^N \omega_k^{(\alpha)} \phi_i^{(\alpha)}(x_k^{(\alpha)}) \phi_j^{(\alpha)}(x_k^{(\alpha)}) = \delta_{ij}, \quad (4.32)$$

where  $x_k^{(\alpha)} \in S_N^{(\alpha)}$ . Assume that in iteration  $m$ , the solution function  $y(x)$  is approximated by

$$y(x) \approx \sum_{i=0}^N a_i C_i^{(\alpha_m)}(x). \quad (4.33)$$

## Chapter 4

---

The Gegenbauer collocation coefficients  $\{a_i\}_{i=0}^N$  are computed by interpolating the solution function  $y(x)$  at the elements of  $S_N^{(\alpha_0)}$  such that

$$y(x_k^{(\alpha_0)}) = \sum_{i=0}^N a_i C_i^{(\alpha_m)}(x_k^{(\alpha_0)}). \quad (4.34)$$

Multiplying both sides by  $\omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)})$ , and summing over  $k$  yields

$$\sum_{k=0}^N \omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)}) y(x_k^{(\alpha_0)}) = \sum_{i=0}^N a_i \sum_{k=0}^N \omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)}) C_i^{(\alpha_m)}(x_k^{(\alpha_0)}) \quad (4.35)$$

$$= \sum_{i=0}^N a_i (\lambda_j^{(\alpha_m)} \lambda_i^{(\alpha_m)})^{\frac{1}{2}} \sum_{k=0}^N \omega_k^{(\alpha_m)} \phi_j^{(\alpha_m)}(x_k^{(\alpha_0)}) \phi_i^{(\alpha_m)}(x_k^{(\alpha_0)}) \quad (4.36)$$

$$\neq \sum_{i=0}^N a_i (\lambda_j^{(\alpha_m)} \lambda_i^{(\alpha_m)})^{\frac{1}{2}} \delta_{ji} = a_j \lambda_j^{(\alpha_m)}, \quad (4.37)$$

since the discrete orthonormality relation (4.32) is not satisfied. Hence

$$a_j \neq (\lambda_j^{(\alpha_m)})^{-1} \sum_{k=0}^N \omega_k^{(\alpha_m)} C_j^{(\alpha_m)}(x_k^{(\alpha_0)}) y(x_k^{(\alpha_0)}) \quad \forall j, \quad (4.38)$$

and Equation (4.30) is false. In fact, the coefficients  $\{a_j\}_{j=0}^N$  cannot be obtained by Equation (4.30) via interpolation at the elements of  $S_N^{(\alpha_0)}$  for any value of  $\alpha_0$ , since the interpolation/collocation nodes  $x_k^{(\alpha_0)} \in S_N^{(\alpha_0)}$  used in the approximation are the zeros of the Gegenbauer polynomial  $C_{N+1}^{(\alpha_0)}(x)$ , while we can clearly see that the Gegenbauer polynomials occurring in Equation (4.30) are associated with a different Gegenbauer parameter  $\alpha_m$ . Hence, the form of the Gegenbauer integration matrix elements defined by Equation (4.27) is inconsistent to approximate the integral operations of the unknown solution  $y(x)$  at the elements of  $S_N^{(\alpha_0)}$ . Therefore optimizing the Gegenbauer integration matrix through an equivalent unconstrained optimization problem with  $\alpha$  added as an extra optimization variable violates the discrete Gegenbauer orthonormality relation; consequently, the numerical scheme may produce false solution approximations.

## 4.5 Concluding Remarks and a Practical Alternative Method

The discrete orthonormality relation (4.32) is valid with respect to the measure  $(1-x^2)^{\alpha-1/2}$  in the space  $L^2([-1, 1])$ , and as the parameter  $\alpha$  changes, the pro-

## Chapter 4

---

jection space  $L^2([-1, 1])$  changes. Since the eigenfunctions in spectral theory are held fixed to define the projection space, and the approximation is carried out with respect to that space, the approximation procedure must start anew as the space is refined. Hence one cannot change the projection space while performing the approximation; consequently, one cannot solve a mathematical problem correctly by optimizing the Gegenbauer integration matrix through the addition of the Gegenbauer parameter  $\alpha$  in the equivalent unconstrained/constrained optimization problem.

To avoid the violation of the discrete Gegenbauer orthonormality relation, the collocation nodes  $\{x_k^{(\alpha_0)}\}_{k=0}^N$  must be replaced with the elements of  $S_N^{(\alpha_m)}$ , which is impossible since the  $x$ 's represent the solution nodes of the problem, and they must be fixed during the whole optimization procedure. A different path to overcome the problem is to fix the values of the integration nodes  $\{x_i^{(\alpha_0)}\}_{i=0}^N$  representing the upper limit of the integration in Equation (4.27) while varying the values of the other discretization nodes  $\{x_k^{(\alpha_0)}\}_{k=0}^N$  in the equation into  $\{x_k^{(\alpha_m)}\}_{k=0}^N$  to be consistent with the corresponding value of  $\alpha_m$ . Yet, we still have the problem of evaluating the unknown solution function  $y(x)$  at the nodes  $\{x_k^{(\alpha_m)}\}_{k=0}^N$  in order to perform the integration given by Equation (4.26). The latter cannot be established without another means of approximation such as interpolation after obtaining some certain values of the unknown solution function  $y(x)$  using any integral equation solvers, questioning the accuracy and the efficiency of the Gegenbauer collocation method. Notice here that the sought accuracy obtained using the Gegenbauer collocation method cannot exceed the precision of the approximated values  $\{y(x_k^{(\alpha_m)})\}_{k=0}^N$  obtained from the interpolation method. Hence optimizing the Gegenbauer integration matrix by adding the Gegenbauer parameter  $\alpha$  as an extra unknown variable in the resulting optimization problem, and then optimizing  $\alpha$  via an algorithmic procedure starting with an arbitrary choice  $\alpha_0$  is deceptive. This method cannot work properly unless the value of  $\alpha_m$  is very close from the initial value  $\alpha_0$  at each iteration, so that the values of the discretization nodes  $\{x_k^{(\alpha_m)}\}_{k=0}^N$  and the solution nodes  $\{x_i^{(\alpha_0)}\}_{i=0}^N$  become very close, and Equations (4.26) and (4.27) may still provide reasonable approximations to the integrations of the unknown solution function  $y(x)$  at the solution nodes  $\{x_i^{(\alpha_0)}\}_{i=0}^N$ .

We draw the attention of the reader that in (Elgindy and Smith-Miles, 2013b) we have presented a practical alternative method for optimizing the Gegenbauer integration matrix. In particular, we have introduced the idea of optimally constructing the Gegenbauer quadrature through discretizations at some optimal sets of Gegenbauer-Gauss points in a certain optimality sense. We showed that the Gegenbauer polynomial expansions can produce higher-order approximations to the definite integrals  $\int_{-1}^{x_i} f(x)dx, i = 0, \dots, N$  of some function  $f(x) \in C^\infty[-1, 1]$

## Chapter 4

---

for the small/medium range of the number of Gegenbauer expansion terms by minimizing the quadrature error at each integration point  $x_i$  through a pointwise approach. This technique entails the calculation of some optimal Gegenbauer parameter values  $\{\alpha^{(i)}\}_{i=0}^N$  rather than choosing any arbitrary  $\alpha$  value. This key idea has been applied later in (Elgindy and Smith-Miles, 2013a,c; Elgindy et al., 2012) for the solutions of boundary-value problems, integral equations, integro-differential equations, and optimal control problems. The presented methods allow for optimizing the Gegenbauer parameter  $\alpha$  internally during the construction of the Gegenbauer integration matrix to achieve higher-order approximations without adding it as an extra unknown variable after converting the problem into an optimization one. The results clearly showed that determining an optimal set of the Gegenbauer parameters  $\{\alpha^{(i)}\}_{i=0}^N$  in a certain optimality sense allow the Gegenbauer integration matrices to produce higher-order approximations which can exceed those obtained by the standard Chebyshev, Legendre, and Gegenbauer integration matrices at each solution node  $\{x_i\}_{i=0}^N$  at least for a small number of spectral expansion terms. The interested reader may consult (Elgindy and Smith-Miles, 2013a,b,c; Elgindy et al., 2012) for more information on the developed numerical methods.

This page is intentionally left blank

## PART B: Suggested Declaration for Thesis Chapter

Monash University

### Declaration for Thesis Chapter 5

#### Declaration by candidate

In the case of Chapter 5, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
The author of the key ideas, programming codes, organization, development, and writing up of the article	90

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided valuable comments and aided proofreading	

Candidate's  
Signature

	Date 11/05/2013
--	--------------------

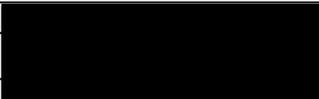
#### Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)

Signature 1

	Date 14/5/13
---	--------------

.....

This page is intentionally left blank

## *Chapter 5*

# *Solving Boundary Value Problems, Integral, and Integro-Differential Equations Using Gegenbauer Integration Matrices*

**Chapter 5 is based on the published article Elgindy, K. T., Smith-Miles, K. A., 1 January 2013. Solving boundary value problems, integral, and integro-differential equations using Gegenbauer integration matrices. *Journal of Computational and Applied Mathematics* 237 (1), 307–325.**

**Abstract.** *We introduce a hybrid Gegenbauer (ultraspherical) integration method (HGIM) for solving boundary value problems (BVPs), integral and integro-differential equations. The proposed approach recasts the original problems into their integral formulations, which are then discretized into linear systems of algebraic equations using Gegenbauer integration matrices (GIMs). The resulting linear systems are well-conditioned and can be easily solved using standard linear system solvers. A study on the error bounds of the proposed method is presented, and the spectral convergence is proven for two-point BVPs (TPBVPs). Comparisons with other competitive methods in the recent literature are included. The proposed method results in an efficient algorithm, and spectral accuracy is verified using eight test examples addressing the aforementioned classes of problems. The proposed method can be applied on a broad range of mathematical problems while producing highly accurate results. The developed numerical scheme provides a viable alternative to other solution methods when high-order approximations are required using only a relatively small number of solution nodes.*

**Keyword.** *Collocation points; Gegenbauer-Gauss points; Gegenbauer integration matrix; Gegenbauer integration method; Gegenbauer polynomials; Spectral methods.*

References are considered at the end of the thesis.

## Chapter 5

# Solving Boundary Value Problems, Integral, and Integro-Differential Equations Using Gegenbauer Integration Matrices

### 5.1 Introduction

A number of physical phenomena can be modeled as ODEs, integral equations, or integro-differential equations. Spectral methods have been one of the most elegant methods by far for solving these problems. They provide a computational approach which has achieved substantial popularity over the past three decades, and has been widely used for the numerical solutions of various differential and integral equations; see (Bernardi and Maday, 1997; Boyd, 1989; Canuto et al., 1988, 2006; Delves and Mohamed, 1985; Funaro, 1992; Gottlieb and Orszag, 1977; Guo, 1998; Tian, 1989). In this class of methods, the approximation of functions in  $C^\infty[a, b]$  can be performed using truncated series of the eigenfunctions of certain singular Sturm-Liouville problems. It is well-known that the truncation error approaches zero faster than any negative power of the number of basis functions used in the approximation, as that number (order of truncation  $N$ ) tends to infinity (Hosseini, 2006). The latter phenomenon is usually referred to as the “spectral accuracy” (Gottlieb and Orszag, 1977). The principal advantage of spectral methods lies in their ability to achieve accurate results with substantially fewer degrees of freedom.

For differential equations, spectral methods transform the problems into al-

## Chapter 5

---

gebraic linear systems of equations via approximating the unknown solution by a truncated spectral expansion series and its derivatives by spectral differentiation matrices (SDMs). The latter are linear maps which take a vector of  $N$  function values  $f(x_i)$  to a vector of  $N$  derivative values  $f'(x_i)$ . This is extraordinarily accurate in exact arithmetic; however there are a number of difficulties associated with the practical implementation as SDMs are known to be severely ill-conditioned (Funaro, 1987); the condition number of the  $N^{\text{th}}$ -order differentiation matrix scales best as  $O(N^{2k})$ , where  $k$  is the order of the derivative of the solution function (Hesthaven, 2000). Consequently the ill-conditioning of SDMs with increasing order frequently causes degradation of the observed precision (Greengard, 1991; Tang and Trummer, 1996) as the procedure involves the solution of very ill-conditioned linear system of equations, and the need for developing efficient preconditioners becomes crucial (Elbarbary, 2006; Hesthaven, 1998). Trefethen (1988); Trefethen and Trummer (1987) showed that the time step restrictions due to this ill-conditioning can be more severe than those predicted by the standard stability theory.

Another approach for solving differential equations is to recast the governing differential equation as an integral equation, and then discretize the latter using spectral integration matrices (SIMs) into an algebraic linear system of equations, which is then solved with spectral accuracy. A SIM can be defined similarly to a SDM as a linear map which takes a vector of  $N$  function values  $f(x_i)$  to a vector of  $N$  integral values  $\int_a^{x_i} f(x)dx$ , for some real number  $a \in \mathbb{R}$ . This strategy eludes the drawbacks of applying SDMs as SIMs are known to be well-conditioned operators (Elbarbary, 2006, 2007; Elgindy, 2009; Greengard, 1991; Lundbladh et al., 1992); their well-conditioning is essentially unaffected for increasing number of points (Elgindy, 2009). Moreover, the use of integration operations for constructing spectral approximations improves their rate of convergence, and allows the multiple boundary conditions to be incorporated more efficiently (Elgindy, 2009; Mai-Duy and Tanner, 2007). In fact, the application of integration operators for the treatment of differential equations by orthogonal polynomials, in particular, Chebyshev polynomials, dates back to Clenshaw (1957) in the late 1950's. The spectral approximation of the integration form of differential equations was put forward later in the 1960's by Clenshaw and Curtis in the spectral space and by El-Gendi (1969) in the physical space. For an  $N^{\text{th}}$ -order differential equation, the approximate solutions are obtained by either recasting the  $N^{\text{th}}$ -order differential equation directly into its integral form for the solution function, or by using the indefinite integral and spectral quadratures to transform the differential equation into its integral form for the  $N^{\text{th}}$ -order derivative. The solution and its derivatives up to the  $(N - 1)^{\text{th}}$ -order derivative may then be stably recovered by integration. The latter idea was brought forward by Greengard in 1991, and was applied successfully on TPBVPs using Clenshaw-Curtis quadrature (Clenshaw

## Chapter 5

---

and Curtis, 1960). This strategy developed by Greengard (1991) was described later by Driscoll et al. (2008) as a “powerful idea,” since it avoids the matter of loss of digits in spectral methods related to the ill-conditioning of the associated matrices.

The reason for the success of the spectral integration approaches is basically because differentiation is inherently sensitive, as small perturbations in data can cause large changes in result, while integration is inherently stable. Moreover, the integral equation formulation, when applied to TPBVPs for instance, is insensitive to boundary layers, insensitive to end-point singularities, and leads to small condition numbers while achieving high computational efficiency (Greengard and Rokhlin, 1991). The aforementioned spectral integration methods were widely accepted and applied by many authors. Coutsias et al. (1996a,b) extended Greengard’s technique to more general problems, and proved the boundedness of the condition numbers as the discretization is refined (Driscoll, 2010). Mihaila and Mihaila (2002) presented a Chebyshev numerical scheme based on El-Gendi’s method (El-Gendi, 1969) for solving initial value problems and second-order BVPs defined on finite domains. Their method recasts the differential equation into an algebraic equations system which is then solved directly for the values of the solution function at the zeros/extrema of the  $N^{\text{th}}$ -degree Chebyshev polynomial. Elbarbary (2007) extended El-Gendi’s method (El-Gendi, 1969) and developed a spectral successive integration matrix based on Chebyshev expansions for the solution of BVPs after recasting the latter into integral equations. Mai-Duy et al. (2008) reported a global fictitious-domain/integral-collocation method for the numerical solution of second-order elliptic PDEs in irregularly shaped domains, where the construction of the Chebyshev approximations representing the dependent variable and its derivatives are based on integration rather than conventional differentiation. Later, Elgindy (2009) extended El-Gendi’s method (El-Gendi, 1969) and developed some higher order pseudospectral integration matrices based on Chebyshev polynomials, and applied the method on some initial value problems, BVPs, linear integral and integro-differential equations. Driscoll (2010) generalized Greengard’s method (Greengard, 1991) to  $m^{\text{th}}$ -order boundary value and generalized eigenvalue problems, where large condition numbers associated with differentiation matrices in high-order problems are avoided. The area of integro-differential equations can be treated using similar ideas applied to differential equations. It is noteworthy to mention that the solution of integral equations may be obtained analytically using the theory developed by Muskhelishvili (1953), and many different methods for solving integral equations analytically are described in several books; see (Dzhuraev, 1992; Green, 1969; Kanwal, 1997). Furthermore, there are extensive works in the literature for solving integral equations using well-developed numerical integration tools; see (Atkinson, 1997; Delves and Mohamed, 1985; Golberg, 1990). However, if the solutions of the

## Chapter 5

---

integral equations are sufficiently smooth, then it is necessary to consider “very high-order numerical methods” such as spectral methods for approximating the solutions (Tang et al., 2008).

To solve integral equations numerically using spectral methods, definite integrations involved in the equations are approximated using SIMs. The construction of the latter depends on the choice of the nodes  $\{x_i\}_{i=0}^N$  and the basis functions of the approximating series. Recently, in (Elgindy and Smith-Miles, 2013b), we have developed an optimal GIM quadrature (P-matrix quadrature) for approximating definite integrations using Gegenbauer polynomials. The numerical experiments shown in (Elgindy and Smith-Miles, 2013b) suggest that the Gegenbauer polynomials as basis polynomials can perform better than their popular subclasses, Chebyshev and Legendre polynomials, for approximating definite integrations at least for a small number of the spectral expansion terms. The developed P-matrix possesses several advantages such as it can be applied for approximating integrals at any arbitrary sets of integration nodes, while maintaining higher-order approximations. Also, higher-order approximations can be achieved by increasing the number of its columns without the need to increase the number of the integration nodes. Moreover, the construction of the developed integration matrix is induced by the set of the integration nodes regardless of the integrand function. Gegenbauer polynomials have been applied extensively in many research areas, and have been demonstrated to provide excellent approximations to analytic functions; see (Archibald et al., 2003; Barrio, 1999; Doha and Abd-Elhameed, 2009; Elgindy and Smith-Miles, 2013b; Gelb, 2004; Gottlieb and Shu, 1995b; Lurati, 2007; Malek and Phillips, 1995; Phillips and Karageorghis, 1990; Vozovoi et al., 1996, 1997; Yilmazer and Kocar, 2008).

In this chapter, we shall generalize the path paved by Clenshaw and Curtis (1960), El-Gendi (1969), and Greengard (1991) by recasting the governing differential/integro-differential equations into their integral reformulations, and investigate the application of the recently developed P-matrix quadrature for the solution of the problem. Our focus will be on developing an efficient Gegenbauer integration method, with the concrete aim of comparing it with other available numerical methods in the literature for solving BVPs, integral and integro-differential equations. The proposed approach involves using the P-matrix quadrature together with the  $\hat{P}$ -matrix quadrature (Elgindy and Smith-Miles, 2013b)– a modified form of the GIM quadrature developed earlier by El-Hawary et al. (2000)– for recasting the integral equations into a system of linear equations. MATLAB 7 linear system solver is then implemented for solving the resulting algebraic linear system, where accurate results can be obtained. We provide a rigorous error analysis of the proposed method for TPBVPs, which indicates that the numerical errors decay exponentially provided that the unknown and variable coefficient functions are sufficiently smooth. The remaining part of this chapter

## Chapter 5

---

is organized as follows: In section 2 we introduce the Gegenbauer polynomials and some of their properties, with a special attention being given to the construction of the GIMs. In the following section, a discussion on the solution of a linear TPBVP is presented using GIMs. Section 5.2.2 is devoted to a study on the convergence rate and error estimation of the proposed method. The performance of the proposed method is illustrated in Section 5.3, with several practical examples on BVPs, integral and integro-differential equations demonstrating the efficiency and accuracy of our proposed method as well as its generality. Finally, in Section 5.4, we provide some concluding remarks summarizing the advantages of the proposed approach.

### 5.2 The GIMs

We shall require several results from approximation theory before presenting the proposed method in Section 5.2.1. The Gegenbauer polynomials  $C_n^{(\alpha)}(x)$ ,  $n \in \mathbb{Z}^+$ , associated with the real parameter  $\alpha > -1/2$ , appear as eigensolutions to the singular Sturm-Liouville problem in the finite domain  $[-1, 1]$ , with the first two being

$$C_0^{(\alpha)}(x) = 1, \quad C_1^{(\alpha)}(x) = 2\alpha x,$$

while the remaining polynomials are given through the recursion formula

$$(n+1)C_{n+1}^{(\alpha)}(x) = 2(n+\alpha)x C_n^{(\alpha)}(x) - (n+2\alpha-1)C_{n-1}^{(\alpha)}(x).$$

The weight function for the Gegenbauer polynomials is the even function  $(1-x^2)^{\alpha-1/2}$ . The Gegenbauer polynomials satisfy the orthogonality relation (Szegő, 1975)

$$\int_{-1}^1 (1-x^2)^{\alpha-1/2} C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = h_n^{(\alpha)} \delta_{mn}, \quad (5.1)$$

where

$$h_n^{(\alpha)} = \frac{2^{1-2\alpha} \pi \Gamma(n+2\alpha)}{n!(n+\alpha)\Gamma^2(\alpha)}; \quad (5.2)$$

$\delta_{mn}$  is the Kronecker delta function. In a typical Gegenbauer spectral method, the unknown solution  $y$  is expanded as a finite series of the Gegenbauer basis polynomials  $C_k^{(\alpha)}(x)$  in the form

$$y(x) \approx \sum_{k=0}^N a_k C_k^{(\alpha)}(x), \quad (5.3)$$

where  $a_k$  are the Gegenbauer spectral expansion coefficients of the solution. For infinitely differentiable solution functions, the produced approximation error,

## Chapter 5

---

when  $N$  tends to infinity, approaches zero with exponential rate. Doha (1991) standardized the Gegenbauer polynomials so that

$$C_n^{(\alpha)}(1) = 1, \quad n = 0, 1, 2, \dots \quad (5.4)$$

As a result of this standardization,  $C_n^{(0)}(x)$  becomes identical with the Chebyshev polynomials of the first kind  $T_n(x)$ ,  $C_n^{(1/2)}(x)$  is the Legendre polynomial  $L_n(x)$ ;  $C_n^{(1)}(x)$  is equal to  $(1/(n+1))U_n(x)$ , where  $U_n(x)$  is the Chebyshev polynomial of the second type (Doha, 1991). Moreover, the Gegenbauer polynomials can be generated by Rodrigues' formula in the following form:

$$C_n^{(\alpha)}(x) = \left(-\frac{1}{2}\right)^n \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} (1-x^2)^{\frac{1}{2}-\alpha} \frac{d^n}{dx^n} (1-x^2)^{n+\alpha-\frac{1}{2}}, \quad (5.5)$$

or starting from

$$C_0^{(\alpha)}(x) = 1, \quad (5.6a)$$

$$C_1^{(\alpha)}(x) = x, \quad (5.6b)$$

the Gegenbauer polynomials can be generated using the following useful recurrence equation:

$$(j+2\alpha)C_{j+1}^{(\alpha)}(x) = 2(j+\alpha)x C_j^{(\alpha)}(x) - j C_{j-1}^{(\alpha)}(x), \quad j \geq 1. \quad (5.6c)$$

The Gegenbauer polynomials satisfy the orthogonality relation (Elgindy and Smith-Miles, 2013b)

$$\int_{-1}^1 (1-x^2)^{\alpha-\frac{1}{2}} C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = \lambda_n^{(\alpha)} \delta_{mn}, \quad (5.7)$$

where

$$\lambda_n^{(\alpha)} = \frac{2^{2\alpha-1} n! \Gamma^2(\alpha + \frac{1}{2})}{(n+\alpha) \Gamma(n+2\alpha)}. \quad (5.8)$$

For the rest of the chapter, by the Gegenbauer polynomials, we shall refer to the Gegenbauer polynomials standardized so that Eq. (5.4) is satisfied. Moreover, by the Chebyshev polynomials, we shall refer to the Chebyshev polynomials of the first kind  $T_n(x)$ . The following theorem gives the truncation error of approximating a smooth function using the Gegenbauer expansion series:

**Theorem 5.2.1** (Truncation error). *Let  $y(x) \in C^\infty[-1, 1]$  be approximated by the Gegenbauer expansion series (5.3), then for each  $x \in [-1, 1]$ , a number  $\xi(x) \in [-1, 1]$  exists such that the truncation error  $E_T(x, \xi, N, \alpha)$  is given by*

$$E_T(x, \xi, N, \alpha) = \frac{y^{(N+1)}(\xi)}{(N+1)! K_{N+1}^{(\alpha)}} C_{N+1}^{(\alpha)}(x), \quad (5.9)$$

## Chapter 5

---

where

$$K_{N+1}^{(\alpha)} = 2^N \frac{\Gamma(N + \alpha + 1)\Gamma(2\alpha + 1)}{\Gamma(N + 2\alpha + 1)\Gamma(\alpha + 1)}. \quad (5.10)$$

*Proof.* See (El-Hawary et al., 2000).  $\square$

One approach for constructing the entries of the GIM was introduced by El-Hawary et al. (2000), and modified later by Elgindy and Smith-Miles (2013b) in the following theorem:

**Theorem 5.2.2.** *Let*

$$S_N^{(\alpha)} = \{x_k | C_{N+1}^{(\alpha)}(x_k) = 0, k = 0, \dots, N\}, \quad (5.11)$$

be the set of the Gegenbauer-Gauss (GG) nodes;  $f(x) \in C^\infty[-1, 1]$  be approximated by the Gegenbauer polynomials, then there exists a matrix  $\hat{P}^{(1)} = (\hat{p}_{ij}^{(1)})$ ,  $i, j = 0, \dots, N$ ; some numbers  $\xi_i \in [-1, 1]$  satisfying

$$\int_{-1}^{x_i} f(x) dx = \sum_{k=0}^N \hat{p}_{ik}^{(1)}(\alpha) f(x_k) + E_N^{(\alpha)}(x_i, \xi_i), \quad (5.12)$$

where

$$\hat{p}_{ik}^{(1)}(\alpha) = \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} \omega_k^{(\alpha)} C_j^{(\alpha)}(x_k) \int_{-1}^{x_i} C_j^{(\alpha)}(x) dx, \quad (5.13)$$

$$(\omega_k^{(\alpha)})^{-1} = \sum_{j=0}^N (\lambda_j^{(\alpha)})^{-1} (C_j^{(\alpha)}(x_k))^2, x_k \in S_N^{(\alpha)}, \quad (5.14)$$

$$\lambda_j^{(\alpha)} = \frac{2^{2\alpha-1} j! \Gamma^2(\alpha + \frac{1}{2})}{(j + \alpha) \Gamma(j + 2\alpha)}; \quad (5.15)$$

$$E_N^{(\alpha)}(x_i, \xi_i) = \frac{f^{(N+1)}(\xi_i)}{(N+1)! K_{N+1}^{(\alpha)}} \int_{-1}^{x_i} C_{N+1}^{(\alpha)}(x) dx. \quad (5.16)$$

*Proof.* See (Elgindy and Smith-Miles, 2013b).  $\square$

The entries of the  $\hat{P}$ -matrix of order  $n$  are given by

$$\hat{p}_{ij}^{(n)} = \frac{(x_i - x_j)^{n-1}}{(n-1)!} \hat{p}_{ij}^{(1)}, \quad i, j = 0, \dots, N \forall x \in [-1, 1]. \quad (5.17)$$

Moreover

$$\hat{p}_{ij}^{(n)} = \frac{(x_i - x_j)^{n-1}}{2^n (n-1)!} \hat{p}_{ij}^{(1)}, \quad i, j = 0, \dots, N \forall x \in [0, 1]. \quad (5.18)$$

## Chapter 5

---

A different approach for constructing the entries of the GIM was introduced by Elgindy and Smith-Miles (2013b). The developed GIM is referred to as the P-matrix; it has been demonstrated to produce higher-order approximations than the  $\hat{P}$ -matrix, especially by increasing the number of its columns. Our approach for constructing the entries of the P-matrix can be described in the following theorem:

**Theorem 5.2.3.** *Let*

$$S_{N,M} = \{z_{i,k} | C_{M+1}^{(\alpha_i^*)}(z_{i,k}) = 0, i = 0, \dots, N; k = 0, \dots, M\}, \quad (5.19)$$

be the generalized set of the GG nodes, where

$$\alpha_i^* = \underset{\alpha > -1/2}{\operatorname{argmin}} \eta_{i,M}^2(\alpha), \quad (5.20)$$

$$\eta_{i,M}(\alpha) = \int_{-1}^{x_i} C_{M+1}^{(\alpha)}(x) dx / K_{M+1}^{(\alpha)}; \quad (5.21)$$

$$K_{M+1}^{(\alpha)} = 2^M \frac{\Gamma(M + \alpha + 1)\Gamma(2\alpha + 1)}{\Gamma(M + 2\alpha + 1)\Gamma(\alpha + 1)}. \quad (5.22)$$

Moreover, let  $f(x) \in C^\infty[-1, 1]$  be approximated by the Gegenbauer polynomials, then there exists a matrix

$P^{(1)} = (p_{ij}^{(1)}), i = 0, \dots, N; j = 0, \dots, M;$  some numbers  $\xi_i \in [-1, 1]$  satisfying

$$\int_{-1}^{x_i} f(x) dx = \sum_{k=0}^M p_{ik}^{(1)}(\alpha_i^*) f(z_{i,k}) + E_M^{(\alpha_i^*)}(x_i, \xi_i), \quad (5.23)$$

where

$$p_{ik}^{(1)}(\alpha_i^*) = \sum_{j=0}^M (\lambda_j^{(\alpha_i^*)})^{-1} \omega_k^{(\alpha_i^*)} C_j^{(\alpha_i^*)}(z_{i,k}) \int_{-1}^{x_i} C_j^{(\alpha_i^*)}(x) dx, \quad (5.24)$$

$$(\omega_k^{(\alpha_i^*)})^{-1} = \sum_{j=0}^M (\lambda_j^{(\alpha_i^*)})^{-1} (C_j^{(\alpha_i^*)}(z_{i,k}))^2, \quad z_{i,k} \in S_{N,M}, \quad (5.25)$$

$$\lambda_j^{(\alpha_i^*)} = \frac{2^{2\alpha_i^* - 1} j! \Gamma^2(\alpha_i^* + \frac{1}{2})}{(j + \alpha_i^*) \Gamma(j + 2\alpha_i^*)}; \quad (5.26)$$

$$E_M^{(\alpha_i^*)}(x_i, \xi_i) = \frac{f^{(M+1)}(\xi_i)}{(M+1)!} \eta_{i,M}(\alpha_i^*). \quad (5.27)$$

*Proof.* See (Elgindy and Smith-Miles, 2013b). □

## Chapter 5

---

Notice here that the P-matrix quadrature has the ability to approximate definite integrals at any arbitrary sets of integration nodes using some suitable choices of the Gegenbauer parameter  $\alpha$  for minimizing the quadrature error at each node. Moreover, the P-matrix is a rectangular matrix of size  $(N + 1) \times (M + 1)$ , where  $M$  denotes the highest degree of the Gegenbauer polynomial employed in the computation of the integration of the function  $f$ . To describe the approximation of the definite integration of the function  $f$  in matrix form using the P-matrix, let  $P^{(1)} = (P_0^{(1)} P_1^{(1)} \dots P_N^{(1)})^T$ ,  $P_i^{(1)} = (p_{i,0}^{(1)}, p_{i,1}^{(1)}, \dots, p_{i,M}^{(1)})$ ;  $i = 0, \dots, N$ . Let also  $V$  be a matrix of size  $(M + 1) \times (N + 1)$  defined as  $V = (V_0 V_1 \dots V_N)$ ,  $V_i = (f(z_{i,0}), f(z_{i,1}), \dots, f(z_{i,M}))^T$ ,  $i = 0, \dots, N$ ;  $f(z_{ij})$  is the function  $f$  calculated at the GG nodes  $z_{ij} \in S_{N,M}$ . Then the approximations of the definite integrals  $\int_{-1}^{x_i} f(x) dx$  of  $f$  using the P-matrix are given by

$$\left( \int_{-1}^{x_0} f(x) dx, \int_{-1}^{x_1} f(x) dx, \dots, \int_{-1}^{x_N} f(x) dx \right)^T \approx P^{(1)} \circ V^T, \quad (5.28)$$

where  $\circ$  is the Hadamard product with the elements of  $P^{(1)} \circ V^T$  given by

$$(P^{(1)} \circ V^T)_i = P_i^{(1)} \cdot V_i = \sum_{j=0}^M p_{i,j}^{(1)} f(z_{i,j}), \quad i = 0, \dots, N. \quad (5.29)$$

To calculate the operational matrix of the successive integration of the function  $f$ , let

$$I_i^{(n)} = \int_{-1}^{x_i} \int_{-1}^{t_{n-1}} \dots \int_{-1}^{t_2} \int_{-1}^{t_1} f(t_0) dt_0 dt_1 \dots dt_{n-2} dt_{n-1} \quad \forall 0 \leq i \leq N,$$

be the  $n$ -fold definite integral of the function  $f$ . Then

$$(I_0^{(n)}, I_1^{(n)}, \dots, I_N^{(n)})^T \approx P^{(n)} \circ V^T, \quad (5.30)$$

where  $P^{(n)} = (p_{i,j}^{(n)})$  is the P-matrix of order  $n$  with the entries

$$p_{i,j}^{(n)} = \frac{(x_i - z_{i,j})^{n-1}}{(n-1)!} p_{i,j}^{(1)}, \quad i = 0, 1, \dots, N; j = 0, 1, \dots, M \forall x \in [-1, 1]. \quad (5.31)$$

For the integration over the interval  $[0, 1]$ , Eq. (5.31) is replaced with

$$p_{i,j}^{(n)} = \frac{(x_i - z_{i,j})^{n-1}}{2^n (n-1)!} p_{i,j}^{(1)}, \quad i = 0, 1, \dots, N; j = 0, 1, \dots, M. \quad (5.32)$$

For further information on the implementation of the  $\hat{P}$ -matrix quadrature and the P-matrix quadrature, we refer the interested reader to Ref. (Elgindy and

## Chapter 5

---

Smith-Miles, 2013b). Also further properties of the family of Gegenbauer polynomials can be found in (Abramowitz and Stegun, 1965; Andrews et al., 1999; Bayin, 2006). In the following section, we shall discuss the solution of a linear TPBVP, which arise frequently in engineering and scientific applications using the developed GIMs.

### 5.2.1 The Proposed HGIM

Suppose for simplicity, and without loss of generality, that we have the following linear TPBVP:

$$y''(x) = f(x)y'(x) + g(x)y(x) + r(x), \quad 0 \leq x \leq 1, \quad (5.33a)$$

with the Dirichlet boundary conditions

$$y(0) = \beta, \quad y(1) = \gamma. \quad (5.33b)$$

To ensure that the problem has a unique solution, suppose also that  $f(x)$ ,  $g(x)$ ;  $r(x)$  are continuous on  $[0, 1]$ ;  $g(x) > 0$  on  $[0, 1]$  (Burden and Faires, 2000). We seek the solution of this problem at the GG nodes  $x_i \in S_N^{(\alpha)}$ ,  $i = 0, \dots, N$ , since they are quadratically clustered at the ends of the domain and well suited for high-order polynomial approximation (Hesthaven et al., 2007). Direct integration converts the problem into the following integral counterpart:

$$y(x) = \int_0^x \int_0^x ((g(t) - F(t))y(t) + r(t))dt dx + \int_0^x f(x)y(x)dx + (c_1 - \beta f_0)x + c_2, \quad (5.34)$$

where  $F \equiv f'$ ;  $f_0 = f(0)$ . The constants  $c_1$  and  $c_2$  are chosen to satisfy the boundary conditions such that

$$c_1 = \gamma + \beta(f_0 - 1) - \int_0^1 f(x)y(x)dx - \int_0^1 \int_0^x ((g(t) - F(t))y(t) + r(t))dt dx; \quad (5.35)$$

$$c_2 = \beta. \quad (5.36)$$

Let  $x_{N+1} = 1$ , then applying the P-matrix quadratures recasts the integral equation (5.34) into the following algebraic linear system of equations:

$$w_i - \sum_{j=0}^M (p_{ij}^{(2)}((g_{ij} - F_{ij})w_{ij} + r_{ij}) - p_{ij}^{(1)}f_{ij}w_{ij}) + (\beta f_0 - c_1)x_i - \beta = 0, \quad i = 0, \dots, N; \quad (5.37)$$

## Chapter 5

---

the constant  $c_1$  can be approximated as

$$c_1 \approx \gamma + \beta(f_0 - 1) - \sum_{j=0}^M (p_{N+1,j}^{(1)} f_{N+1,j} w_{N+1,j} + p_{N+1,j}^{(2)} ((g_{N+1,j} - F_{N+1,j}) w_{N+1,j} + r_{N+1,j})), \quad (5.38)$$

where  $w = [w_0, w_1, \dots, w_N]^T$ ,  $w_i \approx y(x_i)$ ,  $\bar{w} = (w_{lj})$ ,  $w_{lj} \approx y(z_{lj})$ ,  $g_{lj} = g(z_{lj})$ ,  $F_{lj} = F(z_{lj})$ ,  $r_{lj} = r(z_{lj})$ ,  $z_{lj} \in S_{N+1,M}$ ,  $i = 0, \dots, N$ ;  $l = 0, \dots, N+1$ ;  $j = 0, \dots, M$ . Hence we have  $(N+2)$  equations in  $M(N+2) + 2N + 3$  unknowns. Since the set of solution nodes  $\{x_i\}_{i=0}^N$  is symmetric, and assuming that the number  $N$  is even, we have  $z_{ij} = z_{N-i,j} \forall 0 \leq i \leq N, 0 \leq j \leq M$ , and the linear system is in fact a system of  $N(M+3)/2 + 2M + 3$  unknowns. Although the P-matrix quadrature presented in (Elgindy and Smith-Miles, 2013b) has been demonstrated to produce high-order approximations, the pure implementation of the P-matrix quadrature for approximating the TPBVP leads to an under-determined linear system of equations. To obtain a square system of equations with a unique solution, we propose to apply a hybrid technique using the P-matrix quadrature and the  $\hat{P}$ -matrix quadrature for the solution of the TPBVP (5.33). The term  $r(x)$  which does not include the unknown function  $y(x)$  will be integrated using the P-matrix quadrature, while the rest of the integrations will be approximated using the  $\hat{P}$ -matrix quadrature. Hence Eqs. (5.37) and (5.38) are replaced with the following two equations:

$$w_i - \sum_{j=0}^N (\hat{p}_{ij}^{(2)} (g_j - F_j) + \hat{p}_{ij}^{(1)} f_j) w_j - \sum_{j=0}^M p_{ij}^{(2)} r_{ij} + (\beta f_0 - c_1) x_i - \beta = 0, \quad i = 0, \dots, N, \quad (5.39)$$

$$c_1 \approx \gamma + \beta(f_0 - 1) - \sum_{j=0}^N (\hat{p}_{N+1,j}^{(1)} f_j + \hat{p}_{N+1,j}^{(2)} (g_j - F_j)) w_j - \sum_{j=0}^M p_{N+1,j}^{(2)} r_{N+1,j}; \quad (5.40)$$

the TPBVP (5.33) is transformed into  $(N+2)$  linear system of algebraic equations in  $(N+2)$  unknowns, which can be written further in the matrix form  $Aw = b$ , where the entries of the coefficient matrix  $A = (a_{ij})$ , and the column vector  $b = (b_i)$  are given by

$$a_{ij} = \delta_{ij} - (\hat{p}_{ij}^{(1)} - \hat{p}_{N+1,j}^{(1)} x_i) f_j + (\hat{p}_{ij}^{(2)} - \hat{p}_{N+1,j}^{(2)} x_i) (F_j - g_j), \quad (5.41a)$$

$$b_i = \sum_{j=0}^M p_{ij}^{(2)} r_{ij} - x_i \left( \sum_{j=0}^M p_{N+1,j}^{(2)} r_{N+1,j} + \beta - \gamma \right) + \beta; \quad i, j = 0, \dots, N. \quad (5.41b)$$

The approximate solutions are then obtained using efficient linear system solvers. One of the advantages of this formulation is that the linear system which arises

## Chapter 5

---

from the discretization is generally well-conditioned (Elbarbary, 2006, 2007; Elgindy, 2009; Greengard and Rokhlin, 1991; Lundbladh et al., 1992). The area of integral equations can be approached directly by the GIMs without any additional reformulations, while similar ideas to the present method can be easily generalized for solving general linear BVPs and integro-differential equations by recasting the original problem into its integral form. The latter can be written generally as

$$Ly = \left( \sum_{j=0}^s f_j(x) I_j \right) y = g(x), \quad x \in [0, 1], s \in \mathbb{Z}^+, \quad (5.42)$$

where  $L = \sum_{j=0}^s f_j(x) I_j$  is a linear integral operator,  $\{f_j\}_{j=0}^s$ ;  $g$  are some known real functions of  $x$ ,  $I_j$  denotes the  $j$ -fold integral of  $y$  with respect to  $x$ ;  $y(x) \in C^\infty[0, 1]$  is the unknown solution of the problem approximated using the Gegenbauer expansion series (5.3). Hence the proposed HGIM can be broadly applied on a wide range of mathematical problems. The following section addresses the convergence rate of the proposed method on the TPBVP (5.33).

### 5.2.2 Convergence Analysis and Error Bounds

Our goal in this section is to show that the rate of convergence is exponential through a convergence analysis of the proposed numerical scheme. We shall provide two lemmas of particular interest for the analysis of the error bounds before presenting the main theorem in this section. The following Lemma highlights the bounds on the Gegenbauer polynomials generated by Eqs. (5.6):

**Lemma 5.2.4.** *The maximum value of the Gegenbauer polynomials  $C_N^{(\alpha)}(x)$  generated by Eqs. (5.6) is less than or equal to 1 for all  $\alpha \geq 0$ ;  $N \geq 0$ , and of order  $N^{-\alpha}$  for all  $-1/2 < \alpha < 0$ ;  $N \gg 1$ .*

*Proof.* We demonstrate two different approaches for proving the first part of the lemma. Firstly, using mathematical induction, it is clear that the lemma is true for  $C_0^{(\alpha)}(x)$  and  $C_1^{(\alpha)}(x)$ . Now assume that the lemma is true for  $N = K \in \mathbb{Z}^+$ , and let  $\beta_{K,\alpha} = 2(K + \alpha)/(K + 2\alpha)$ . For  $N = K + 1$ , we have

$$\begin{aligned} C_{K+1}^{(\alpha)} &= \beta_{K,\alpha} x C_K^{(\alpha)}(x) + (1 - \beta_{K,\alpha}) C_{K-1}^{(\alpha)}(x) \\ \Rightarrow \left| C_{K+1}^{(\alpha)} \right| &\leq |\beta_{K,\alpha}| + |1 - \beta_{K,\alpha}| = 1 \quad \forall \alpha \geq 0. \end{aligned} \quad (5.43)$$

A second approach to prove this result can be derived through the relation between the Gegenbauer polynomials  $C_N^{(\alpha)}(x)$  standardized by (5.4), and the Gegenbauer polynomials  $\hat{C}_N^{(\alpha)}(x)$  standardized by Szegö (1975). Indeed, using Eq.

## Chapter 5

---

(4.7.1) in (Szegő, 1975), and Eq. (1) in (Doha, 2002), we can show that

$$C_N^{(\alpha)}(x) = \frac{\hat{C}_N^{(\alpha)}(x)}{\hat{C}_N^{(\alpha)}(1)} \quad \forall x \in [-1, 1], \alpha > -\frac{1}{2}; N \geq 0. \quad (5.44)$$

From Eq. (7.33.1) in (Szegő, 1975), we have

$$\max_{|x| \leq 1} |\hat{C}_N^{(\alpha)}(x)| = \hat{C}_N^{(\alpha)}(1), \quad (5.45)$$

from which Inequality (5.43) results. To prove the second part of the lemma, we have through Eqs. (7.33.2) and (7.33.3) in (Szegő, 1975) that

$$\max_{|x| \leq 1} |\hat{C}_N^{(\alpha)}(x)| \approx 2^{1-\alpha} |\Gamma(\alpha)|^{-1} N^{\alpha-1} \quad \forall -\frac{1}{2} < \alpha < 0; N \gg 1. \quad (5.46)$$

Since

$$\frac{1}{\hat{C}_N^{(\alpha)}(1)} = \frac{N! \Gamma(2\alpha)}{\Gamma(N+2\alpha)} \approx \frac{\Gamma(2\alpha)}{N^{2\alpha-1}} \quad (\text{asymptotically}), \quad (5.47)$$

then

$$\max_{|x| \leq 1} |C_N^{(\alpha)}(x)| \approx D_\alpha N^{-\alpha} = O(N^{-\alpha}) \text{ as } N \rightarrow \infty, \quad (5.48)$$

where  $D_\alpha$  is a positive constant independent of  $N$ . This completes the proof of the second part of the lemma.  $\square$

The following lemma is substantial for the analysis of the convergence rate of the HGIM:

**Lemma 5.2.5.** *For a fixed  $\alpha > -1/2$ , the factor  $1/((N+1)!K_{N+1}^{(\alpha)})$  is of order  $1/(N^{\frac{1}{2}-\alpha}(2N/e)^N)$ , for large values of  $N$ .*

*Proof.* Using Stirling's formula for the factorial function

$$x! = \sqrt{2\pi x} x^{x+\frac{1}{2}} \exp(-x + \frac{\theta}{12x}), \quad x > 0, 0 < \theta < 1, \quad (5.49)$$

we have

$$\sqrt{2\pi x} x^{x+1/2} e^{-x} < \Gamma(x+1) < \sqrt{2\pi x} x^{x+1/2} e^{-x} e^{\frac{1}{12x}} \quad \forall x \geq 1.$$

Hence

$$\begin{aligned} \frac{\Gamma(N+2\alpha+1)}{\Gamma(N+\alpha+1)} &= \frac{N+\alpha+1}{N+2\alpha+1} \frac{\Gamma(N+2\alpha+2)}{\Gamma(N+\alpha+2)} \\ &< \frac{N+\alpha+1}{N+2\alpha+1} (N+2\alpha+1)^\alpha \left(1 + \frac{\alpha}{N+\alpha+1}\right)^{N+\alpha+\frac{3}{2}} e^{\frac{1}{12(N+2\alpha+1)}-\alpha} \\ &\sim N^\alpha \text{ as } N \rightarrow \infty. \end{aligned}$$

## Chapter 5

---

From Definition (5.10), taking the limit when  $N \rightarrow \infty$  yields

$$1/K_{N+1}^{(\alpha)} \sim 2^{-N} N^\alpha.$$

The proof is established by applying the asymptotic Stirling's approximation formula for the factorial function

$$N! \approx \sqrt{2\pi N} \left(\frac{N}{e}\right)^N \text{ as } N \rightarrow \infty.$$

□

Now, let  $\psi(x) = (g(x) - F(x))y(x)$ ,  $\rho(x) = f(x)y(x)$ , then Eq. (5.34) can be written at the GG collocation points as

$$I_i(y) = y(x_i) - \int_0^{x_i} \int_0^x (\psi(t) + r(t)) dt dx - \int_0^{x_i} \rho(x) dx + (\beta f_0 - c_1)x_i - \beta, \quad i = 0, \dots, N, \quad (5.50)$$

with

$$c_1 = \gamma + \beta(f_0 - 1) - \int_0^{x_{N+1}} \rho(x) dx - \int_0^{x_{N+1}} \int_0^x (\psi(t) + r(t)) dt dx. \quad (5.51)$$

The following theorem highlights the truncation error in approximating the TP-BVP (5.33) using the HGIM:

**Theorem 5.2.6.** *Let the unknown solution  $y(x) \in C^\infty[0, 1]$  of the TPBVP (5.33) be approximated by the Gegenbauer expansion series (5.3), where the Gegenbauer spectral coefficients  $a_k$ 's are calculated by discrete least squares fitting at the GG nodes  $x_i \in S_N^{(\alpha)}$  given by Eq. (5.11). Let  $\tilde{I}_i(w)$  denote the approximation of  $I_i(y)$  using the HGIM for each  $i$ . Also, assume that the functions  $\psi(x)$ ,  $r(x)$ , and  $\rho(x)$  are bounded on the interval  $[0, 1]$  with bounds  $M_1, M_2, M_3$ , respectively. Then for any  $x_i \in S_N^{(\alpha)}$ , there exist some numbers  $\xi_i^{(j)} \in (0, x_i)$ ,  $i = 0, \dots, N+1$ ;  $j = 1, 2, 3$ ;  $\zeta_l \in (0, x_l)$ ,  $l = 0, \dots, N$ , such that*

$$E_T(x_i, \xi_i^{(j)}, \alpha) = I_i(y) - \tilde{I}_i(w) = \frac{y^{(N+2)}(\zeta_i)}{(N+2)!K_{N+2}^{(\alpha)}} C_{N+2}^{(\alpha)}(x_i) + E_{N+1}(x_{N+1}, \xi_{N+1}^{(j)}, \alpha) - E_i(x_i, \xi_i^{(j)}, \alpha), \quad i = 0, \dots, N, \quad (5.52)$$

with

$$\left| E_T(x_i, \xi_i^{(j)}, \alpha) \right| \leq \begin{cases} \hat{D}_i(d_{N+1}^{(\alpha)} |y^{(N+2)}(\zeta_i)| (N+2)^{-\alpha} + (M_1(1+x_i) + 2M_3)d_N^{(\alpha)}(N+1)^{-\alpha} \\ + M_2 d_M^{(\alpha_i^*)} x_i (M+1)^{-\alpha_i^*} + M_2 d_M^{(\alpha_{N+1}^*)} (M+1)^{-\alpha_{N+1}^*}, \\ \quad -1/2 < \alpha < 0, N \gg 1, \\ d_{N+1}^{(\alpha)} |y^{(N+2)}(\zeta_i)| + M_1 d_N^{(\alpha)} (1+x_i) + M_2 (d_M^{(\alpha_{N+1}^*)} + d_M^{(\alpha_i^*)} x_i) \\ + 2M_3 d_N^{(\alpha)}, \quad \alpha \geq 0, \end{cases} \quad (5.53)$$

## Chapter 5

---

where

$$E_i(x_i, \xi_i^{(j)}, \alpha) = \frac{1}{(N+1)!K_{N+1}^{(\alpha)}} \left( \psi^{(N+1)}(\xi_i^{(1)}) \int_0^{x_i} \int_0^x C_{N+1}^{(\alpha)}(t) dt dx \right. \\ \left. + \rho^{(N+1)}(\xi_i^{(3)}) \int_0^{x_i} C_{N+1}^{(\alpha)}(x) dx \right) + \frac{r^{(M+1)}(\xi_i^{(2)})}{(M+1)!K_{M+1}^{(\alpha_i^*)}} \int_0^{x_i} \int_0^x C_{M+1}^{(\alpha_i^*)}(t) dt dx, \quad (5.54)$$

$$d_N^{(\alpha)} = \frac{1}{(N+1)! |K_{N+1}^{(\alpha)}|}, \quad \alpha_i^* = \underset{\alpha > -1/2}{\operatorname{argmin}} \eta_{i,M}^2(\alpha),$$

$$\eta_{i,M}(\alpha) = \int_{-1}^{x_i} C_{M+1}^{(\alpha)}(x) dx / K_{M+1}^{(\alpha)}, \quad i = 0, \dots, N+1,$$

$$\hat{D}_i = \max\{D_\alpha, D_{\alpha_i^*}, D_{\alpha_{N+1}^*}\}, \quad i = 0, \dots, N,$$

$D_\alpha, D_{\alpha_i^*}; D_{\alpha_{N+1}^*}$  are positive constants independent of  $N$ .

*Proof.* The proof can be readily verified. We have

$$\int_0^{x_i} \int_0^x (\psi(t) + r(t)) dt dx + \int_0^{x_i} \rho(x) dx = \chi(x_i, \alpha) + E_i(x_i, \xi_i^{(j)}, \alpha), \quad (5.55)$$

where

$$\chi(x_i, \alpha) = \sum_{j=0}^N (\hat{p}_{ij}^{(2)} \psi_j + \hat{p}_{ij}^{(1)} \rho_j) + \sum_{j=0}^M p_{ij}^{(2)} r_{ij}, \quad i = 0, \dots, N+1;$$

Eq. (5.54) is obtained using formulas (5.16) & (5.27). Hence the approximation to the integral equation (5.50) can be written as

$$\tilde{I}_i(w) = w(x_i) + \chi(x_{N+1}, \alpha) - \chi(x_i, \alpha) - \gamma, \quad i = 0, \dots, N. \quad (5.56)$$

Eq. (5.52) results directly by subtracting Eq. (5.56) from Eq. (5.50), and the truncation error is bounded by

$$\left| E_T(x_i, \xi_i^{(j)}, \alpha) \right| \leq d_{N+1}^{(\alpha)} \left| y^{(N+2)}(\zeta_i) C_{N+2}^{(\alpha)}(x_i) \right| + \left| E_i(x_i, \xi_i^{(j)}, \alpha) \right| + \left| E_{N+1}(x_{N+1}, \xi_{N+1}^{(j)}, \alpha) \right|,$$

where the bounds (5.53) result directly from applying Lemma 5.2.4 and Theorem 5.2.1.  $\square$

## Chapter 5

---

The convergence of the HGIM is illustrated by the decay of the error bounds (5.53), which are mainly affected by the error factor  $d_N^{(\alpha)}$ . While the error bounds might suggest, at the first sight, that the error is smaller for  $\alpha \geq 0$ , Lemma 5.2.5 shows that the error factor  $d_N^{(\alpha)}$  attains its minimum values at the boundary value  $\alpha = -0.5$ . Moreover, Lemma 5.2.5 shows that  $d_N^{(\alpha)}$  is monotonically increasing for increasing values of  $\alpha$ . Consequently applying the HGIM at the GG nodes  $x_i \in S_N^{(\alpha)}$  for negative values of  $\alpha$  seems to be recommended as the truncation error is expected to be smaller, and faster convergence rate is achieved for increasing values of  $N$ . On the other hand, the Gegenbauer polynomials grow rapidly for increasing values of  $N$  as  $\alpha \rightarrow -0.5$ , and only suitable negative values of  $\alpha$  are to be chosen to produce better approximations to the TPBVP (5.33). This error analysis seems to support the numerical experiments conducted by Doha (1990) on the numerical solution of parabolic PDEs, which showed that higher accuracy may be obtained by choosing  $\alpha$  to be “small and negative.” However, our numerical experiments conducted on many test examples in Section 5.3 show that excellent numerical approximations may be obtained through Gegenbauer discretizations for both positive and negative values of  $\alpha$ . Theoretically, the HGIM and Doha (1990)’s Gegenbauer method are different predominantly in the role of the Gegenbauer parameter  $\alpha$  in both methods. Indeed, the HGIM relies on the P-matrix quadrature as part of its numerical integration tools. This in turn indicates that the HGIM produces a truncation error which involves some already determined  $\alpha$  values during the construction of the P-matrix as clearly observed from the error bounds formula (5.53). Consequently, these automatically determined Gegenbauer parameters can greatly affect the magnitude of the truncation error, and excellent numerical approximations may be achieved for Gegenbauer discretizations at both positive and negative values of  $\alpha$ . On the other hand, Doha (1990)’s Gegenbauer method employs Gegenbauer expansion series with a fixed  $\alpha$  value, which favors the negative and small values of  $\alpha$  over the other cases in order to achieve higher precision approximations. The significant result of Lemma 5.2.5 and Theorem 5.2.6 is that the application of the present HGIM for solving a TPBVP, where the unknown solution is assumed to be infinitely differentiable, leads to a spectrally convergent solution. Indeed, the factor  $d_N^{(\alpha)}$ , which appears in each term in the error bounds (5.53), decays exponentially faster than any finite power of  $1/N$ . Figure 5.2.2 confirms this fact, as it shows the logarithm of  $d_N^{(-0.5+\varepsilon)}$  versus different values of  $N$ , where  $\varepsilon = 2.22 \times 10^{-16}$  is the machine epsilon. Another useful result of Lemma 5.2.5 is that increasing the number of the columns of the P-matrix decreases the value of the error factor  $d_M^{(\alpha)}$ , for any choice of  $\alpha > -1/2$ . Hence, the rate of the convergence of the HGIM increases without the need to increase the number of the solution nodes. We have implemented this useful trick on some numerical test examples in Section 5.3, and accomplished

## Chapter 5

---

higher-order approximations (almost full machine precision in some cases) using a relatively small number of solution nodes.

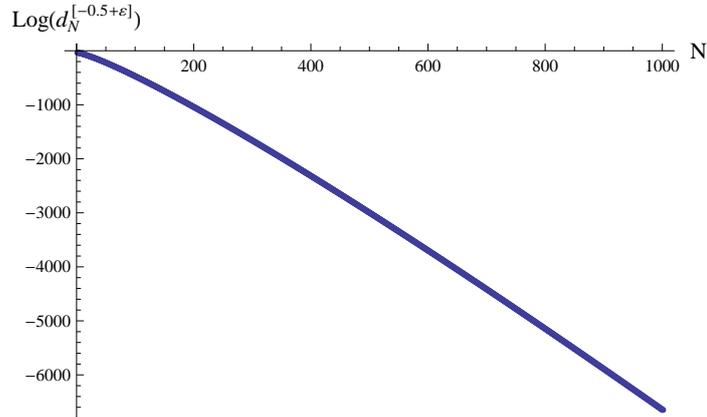


Figure 5.1: The error factor  $d_N^{(\alpha)}$  decays exponentially fast for increasing values of  $N$ .

As a special case, if the terms  $f(x)y'(x)$  and  $g(x)y(x)$  are dropped from the TPBVP (5.33a), then the problem can be solved by a straightforward integration using the P-matrix quadrature at any arbitrary sets of solution nodes.

### 5.3 Numerical Results

In this section eight test examples are solved using the HGIM. The first three test examples are TPBVPs studied by Greengard (1991). The first example was later studied by Greengard and Rokhlin (1991). The fourth and fifth test examples are BVPs studied by Elbarbary (2007) and Zahra (2011), respectively. The sixth test example is a Fredholm integral equation studied by Long et al. (2009). The seventh and eighth test examples are integro-differential equations studied by Maleknejad and Attary (2011) and El-Kady et al. (2009), respectively. All calculations were performed on a personal laptop with a 2.53 GHz Intel Core i5 CPU and 4G memory running MATLAB 7 software in double precision real arithmetic. The solution nodes are the GG nodes  $x_i \in S_N^{(\alpha)}$  defined by Eq. (5.11) for  $\alpha = -0.4 : 0.1 : 1$ ; different values of  $N$ . These choices of the solution nodes are of particular interest since they permit the comparison of the performance of Gegenbauer polynomials with Chebyshev and Legendre polynomials for examples where the  $\hat{P}$ -matrices of different orders are only involved. The P-matrix is constructed via Algorithm 2.2 given in (Elgindy and Smith-Miles, 2013b), with  $M_{\max} = 128$ , and different values of  $M$ . Elgindy and Hedar's line search method

## Chapter 5

---

(Elgindy and Hedar, 2008) is used for determining  $\alpha_i^*$  defined by Eq. (5.20). Here we choose the initial search interval  $[-0.5 + 2\varepsilon, 1]$  based on numerical testing. The line search technique is stopped whenever

$$\left| \frac{d}{d\alpha} \eta_{i,M}^2 \right| < 10^{-16} \wedge \frac{d^2}{d\alpha^2} \eta_{i,M}^2 > 0,$$

is satisfied, where  $\eta_{i,M}$  is the same as defined in Eq. (5.21) for each  $i$ . Hereafter, “MSEs,” and “MAEs,” refer to the observed mean square errors and maximum absolute errors of the present method between the approximations and the exact solutions. The results shown between two round parentheses “(.)” are the value(s) of  $\alpha$  at which the best results are obtained by the present method.

**Example 5.3.1.** Consider the following linear TPBVP:

$$-y'' + 400y = -400\cos^2(\pi x) - 2\pi^2 \cos(2\pi x), \quad y(0) = y(1) = 0, \quad (5.57)$$

with the exact solution

$$y(x) = \frac{e^{-20}}{1 + e^{-20}} e^{20x} + \frac{1}{1 + e^{-20}} e^{-20x} - \cos^2(\pi x).$$

Applying the HGIM recasts the problem into the following algebraic system of linear equations:

$$w_i + 400 \left( \sum_{j=0}^N \hat{p}_{N+1,j}^{(2)} x_i - \hat{p}_{ij}^{(2)} \right) w_j + \sum_{j=0}^M (p_{N+1,j}^{(2)} x_i r_{N+1,j} - p_{ij}^{(2)} r_{ij}) = 0, \quad i = 0, \dots, N, \quad (5.58)$$

where  $r(x) = 400\cos^2(\pi x) + 2\pi^2 \cos(2\pi x)$ . This problem as reported by Stoer and Bulirsch (1980) suffers from the presence of rapidly growing solutions of the corresponding homogeneous equation. In fact, the homogeneous differential equation has solutions of the form  $y(x) = ce^{\pm 20x}$ , which can grow at a rapid exponential rate. Moreover, the derivatives of the exact solution are very large for  $x \approx 0$  and  $x \approx 1$ . The problem was solved by Greengard (1991) using a Chebyshev spectral method (integral equation approach), and Greengard and Rokhlin (1991) by applying a high order Nystrom scheme based on a  $p$ -point Chebyshev quadrature after reducing the differential equation to a second kind integral equation. Comparisons with Greengard and Rokhlin’s method (Greengard and Rokhlin, 1991) are shown in Table 5.1, while comparisons with Greengard’s method (Greengard, 1991) are shown in Table 5.2. Both tables show the greater accuracy obtained by the present HGIM. Moreover, the tables manifest that the Gegenbauer polynomial approximations are very effective in solving TPBVPs for many suitable values of  $\alpha$ . Figure 5.2 shows the numerical behavior of the HGIM, where Figure

## Chapter 5

---

5.2(b) shows the “MSEs” of the present method for  $N = 16, 64$ ;  $M = N$ . Higher-order approximations are obtained via discretizations at positive values of  $\alpha$  for both values of  $N$ . Figure 5.2(c) shows the “MAEs” of the present method for  $N = 7, 15, 23, 31$ . Here again we find that the Gegenbauer discretizations at the positive values of  $\alpha$  are favorable in all four cases.

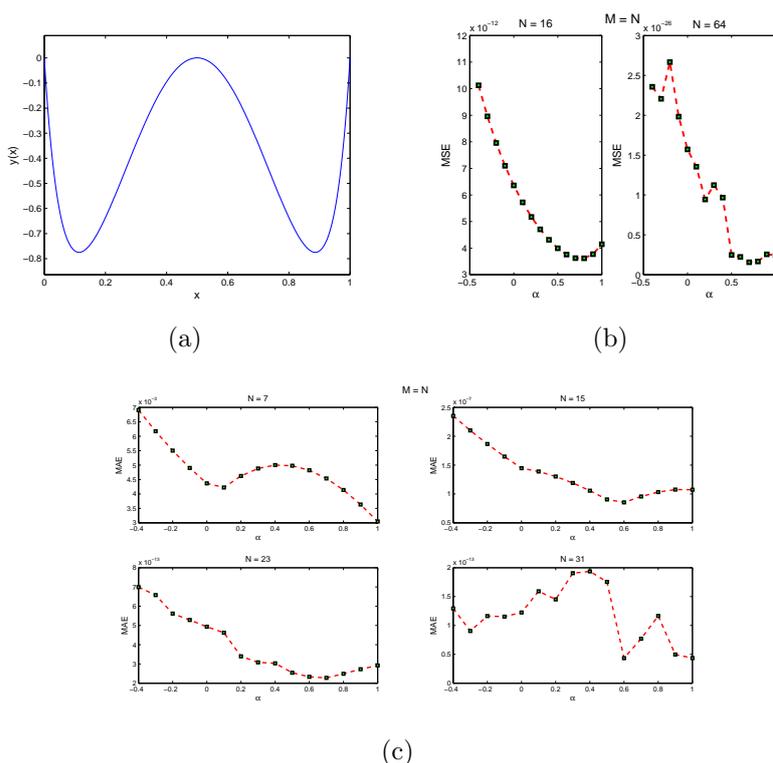


Figure 5.2: The numerical experiments of the HGIM on Example 5.3.1. Figure (a) shows the graph of  $y(x)$  on  $[0, 1]$ . Figure (b) shows the MSEs of the HGIM for  $N = 16; 64$ . Figure (c) shows the MAEs of the HGIM for  $N = 7, 15, 23; 31$ .

**Example 5.3.2.** Consider the following singular perturbation TPBVP:

$$\varepsilon y'' - y = 0, \quad y(-1) = 1, y(1) = 2, \quad (5.59)$$

with  $\varepsilon = 10^{-5}$ .

Setting  $z = (1 + x)/2$  transforms the problem into the following form:

$$4\varepsilon \frac{d^2 y}{dz^2} - y = 0, \quad y(0) = 1, y(1) = 2, \quad (5.60)$$

Example 5.3.1

$N_T$	$(p = 8)$		$(p = 16)$		$(p = 24)$	
	Greengard and Rokhlin (1991)	Present HGIM	Greengard and Rokhlin (1991)	Present HGIM	Greengard and Rokhlin (1991)	Present HGIM
8	$3.16 \times 10^{-02}$	$3.04 \times 10^{-03}$ (1)	16	$6.59 \times 10^{-06}$	24	$5.24 \times 10^{-11}$
16	$1.86 \times 10^{-03}$	$8.54 \times 10^{-08}$ (0.6)	32	$3.01 \times 10^{-09}$		$8.54 \times 10^{-08}$ (0.6)
32	$4.26 \times 10^{-05}$	$4.34 \times 10^{-14}$ (0.6)				$4.34 \times 10^{-14}$ (0.6)
						$2.28 \times 10^{-13}$ (0.7)

Table 5.1: Comparison of the present HGIM with Greengard and Rokhlin's method (Greengard and Rokhlin, 1991).  $N_T$  refers to the total number of nodes. The results shown are the observed MAEs of both methods.

## Chapter 5

---

with the exact solution

$$y(z) = \frac{e^{-50\sqrt{10}z}(-2e^{50\sqrt{10}} + e^{100\sqrt{10}} - e^{100\sqrt{10}z} + 2e^{50\sqrt{10}(2z+1)})}{-1 + e^{100\sqrt{10}}}.$$

The HGIM transforms the problem into the following linear system of equations:

$$4\varepsilon(w_i - z_i - 1) + \sum_{j=0}^N (\hat{p}_{N+1,j}^{(2)} z_i - \hat{p}_{ij}^{(2)}) w_j = 0, \quad i = 0, \dots, N. \quad (5.61)$$

This example represents a clear contest between Legendre, Chebyshev, and Gegenbauer polynomials. It is well-known that Legendre and Chebyshev polynomial expansions give an exceedingly good representation of functions which rapidly change in narrow boundary layers (Gottlieb and Orszag, 1977). Here we show that the Gegenbauer family of polynomials can perform better for several values of  $\alpha$ . Figure 5.3(b) shows the “MSEs” of the present method for  $N = 16, 64; 128$ . The figure shows that the Gegenbauer collocations at the positive values of  $\alpha$  are in favor of the negative values for several values of  $N$ . For large values of  $N$ , collocations at both positive and negative values of  $\alpha$  produce excellent convergence properties. Comparisons between the present method and Greengard’s Chebyshev spectral method (Greengard, 1991) are shown in Table 5.2. The results confirm the spectral decay of the error for increasing values of  $N$ , and the performance of the present method clearly outperforms Greengard’s method (Greengard, 1991). Moreover, the table reveals that the Gegenbauer polynomial approximations are better than those obtained by the Chebyshev and Legendre polynomials for both values of  $N = 16; 64$ .

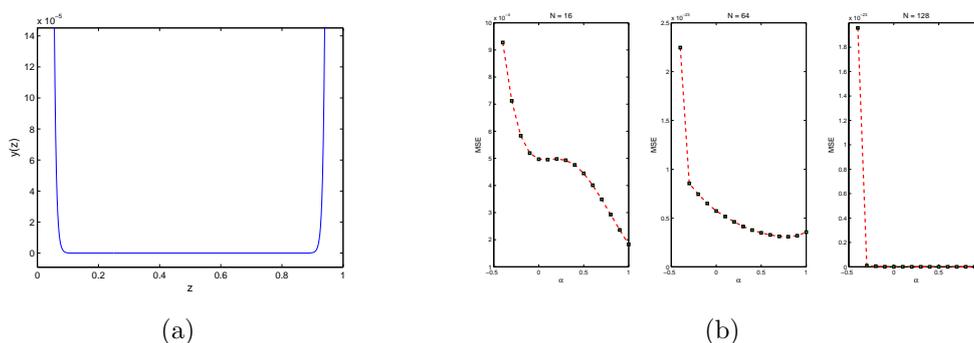


Figure 5.3: The numerical experiments of the HGIM on Example 5.3.2. Figure (a) shows the graph of  $y(z)$  on  $[0, 1]$ . Figure (b) shows the MSEs for  $N = 16, 64; 128$ .

## Chapter 5

---

**Example 5.3.3.** Consider the following linear TPBVP:

$$y'' + 5y' + 10000y = -500 \cos(100x)e^{-5x}, \quad y(0) = 0, y(1) = \sin(100)e^{-5}, \quad (5.62)$$

with the very oscillatory solution  $y(x) = \sin(100x)e^{-5x}$ .

The HGIM transforms the problem into the following linear system of equations:

$$w_i + \sum_{j=0}^N (10000(\hat{p}_{ij}^{(2)} - \hat{p}_{N+1,j}^{(2)}x_i) + 5(\hat{p}_{ij}^{(1)} - \hat{p}_{N+1,j}^{(1)}x_i))w_j + \sum_{j=0}^M (p_{N+1,j}^{(2)}r_{N+1,j}x_i - p_{ij}^{(2)}r_{ij}) - \sin(100)e^{-5}x_i = 0, \quad i = 0, \dots, N. \quad (5.63)$$

Since the solution exhibits oscillatory behavior with ever increasing frequency near the boundary, convergence can only be achieved if sufficient modes are included to resolve the most rapid oscillations present (Coutsias et al., 1996b). Figure 5.4(b) shows that Gegenbauer discretizations at the negative values of  $\alpha$  are generally in favor of the positive values for  $N = 16$ , while discretizations at both positive and negative values of  $\alpha$  share excellent approximation results for  $N = 64$ . Comparisons between the present method and Greengard's method (Greengard, 1991) are shown in Table 5.2. The table suggests that the Gegenbauer polynomials can produce better approximations than Chebyshev and Legendre polynomials.

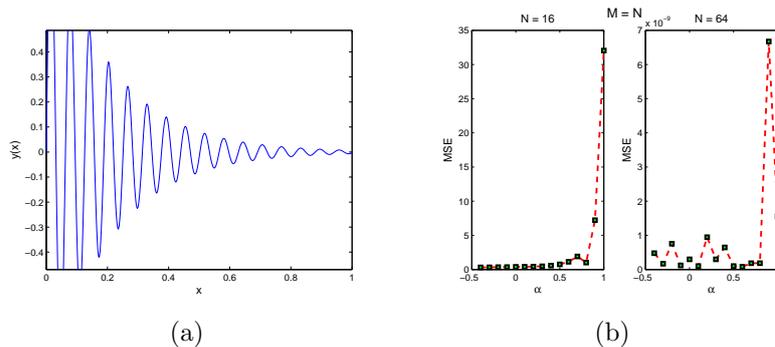


Figure 5.4: The numerical experiments of the HGIM on Example 5.3.3. Figure (a) shows the graph of  $y(x)$  on  $[0, 1]$ . Figure (b) shows the MSEs for  $N = 16; 64$ .

**Example 5.3.4.** Consider the following BVP:

$$y''(x) + xe^{-x}y(x) = e^x + x, \quad -1 \leq x \leq 1, \quad (5.64)$$

$N$	Example 5.3.1		Example 5.3.2		Example 5.3.3	
	Greengard (1991)	Present HGIM	Greengard (1991)	Present HGIM	Greengard (1991)	Present HGIM
16	$7.2 \times 10^{-05}$	$3.6 \times 10^{-12}$ (0.8)	2.2	$1.8 \times 10^{-04}$ (1.0)	0.9	0.3 (-0.4)
64	$8.7 \times 10^{-16}$	$1.6 \times 10^{-27}$ (0.7)	$4.0 \times 10^{-09}$	$3.1 \times 10^{-24}$ (0.8)	$7.9 \times 10^{-04}$	$8.2 \times 10^{-11}$ (0.6)

Table 5.2: Comparison of the present method with Greengard's method (Greengard, 1991). The results shown are the observed MSEs of both methods.

## Chapter 5

---

with the Robin boundary conditions

$$y'(-1) + 2y(-1) = 3e^{-1}, \quad y'(1) - y(1) = 0. \quad (5.65)$$

The problem has the exact solution  $y(x) = e^x$ .

Applying the HGIM yields the following algebraic system of equations:

$$\begin{aligned} w_i + \sum_{j=0}^N (\hat{p}_{i,j}^{(2)} + (1 + 2x_i)(\hat{p}_{N+1,j}^{(1)} - \hat{p}_{N+1,j}^{(2)}))x_j e^{-x_j} w_j \\ - \sum_{j=0}^M (p_{ij}^{(2)} r_{ij} + (p_{N+1,j}^{(2)} - p_{N+1,j}^{(1)})(1 + 2x_i)r_{N+1,j}) + 3e^{-1}x_i = 0, \quad i = 0, \dots, N, \end{aligned} \quad (5.66)$$

where  $r(x) = e^x + x$ . Figure 5.5 shows the numerical behavior of the HGIM using  $M = 16$  for  $N = 8; 10$ . The best approximations for both cases are reported at  $\alpha = 0.6; 1$ . This problem was solved by Elbarbary (2007) using Chebyshev pseudospectral integration matrices after recasting the BVP into its integral form. Comparisons between the HGIM using  $M = 16$  and Elbarbary's method (Elbarbary, 2007) are shown in Table 5.3 for  $N = 8; 10$ . Notice here the rapid convergence accomplished by the present method by increasing the number of the columns of the P-matrix without the need to increase the number of the solution points. Indeed, the table shows that a small number of solution points as  $N = 8$  is sufficient to achieve almost full machine precision. This highly desirable feature is a conspicuous contribution of the HGIM over the standard spectral numerical schemes, and a clear evident on the effectiveness of the Gegenbauer approximation methods over the conventional Chebyshev and Legendre methods.

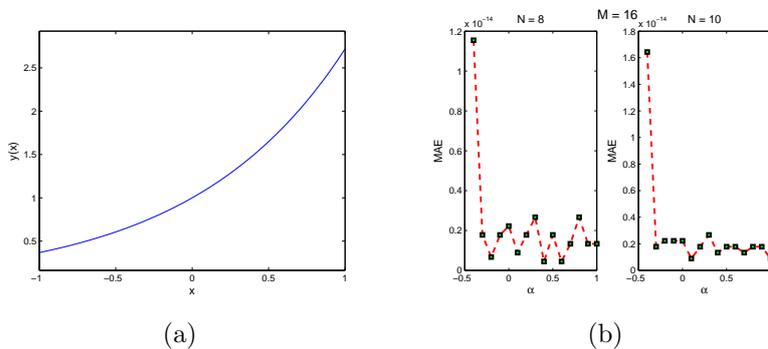


Figure 5.5: The numerical experiments of the HGIM on Example 5.3.4. Figure (a) shows the graph of  $y(x)$  on  $[-1, 1]$ . Figure (b) shows the MAEs for  $N = 8; 10$ .

## Chapter 5

Example 5.3.4				
$x_i$	Elbarbary's Chebyshev pseudospectral integration method (Elbarbary, 2007)	Present HGIM	Elbarbary's Chebyshev pseudospectral integration method (Elbarbary, 2007)	Present HGIM
	$N = 8$		$N = 10$	
$x_0$	$4.0 \times 10^{-10}$	$2.8 \times 10^{-16}$	$4.2 \times 10^{-13}$	$3.9 \times 10^{-16}$
$x_1$	$3.5 \times 10^{-10}$	$1.1 \times 10^{-16}$	$3.9 \times 10^{-13}$	$1.7 \times 10^{-16}$
$x_2$	$4.1 \times 10^{-10}$	$1.1 \times 10^{-16}$	$4.5 \times 10^{-13}$	$2.8 \times 10^{-16}$
$x_3$	$1.2 \times 10^{-10}$	$2.2 \times 10^{-16}$	$1.8 \times 10^{-13}$	0
$x_4$	$5.0 \times 10^{-11}$	0	$2.6 \times 10^{-13}$	0
$x_5$	$9.6 \times 10^{-12}$	$2.2 \times 10^{-16}$	$7.9 \times 10^{-14}$	$1.1 \times 10^{-16}$
$x_6$	$2.8 \times 10^{-10}$	$2.2 \times 10^{-16}$	$1.3 \times 10^{-13}$	$4.4 \times 10^{-16}$
$x_7$	$2.8 \times 10^{-10}$	$4.4 \times 10^{-16}$	$5.9 \times 10^{-13}$	0
$x_8$	$3.3 \times 10^{-10}$	0	$3.3 \times 10^{-13}$	$8.9 \times 10^{-16}$
$x_9$			$3.1 \times 10^{-13}$	$4.4 \times 10^{-17}$
$x_{10}$			$3.5 \times 10^{-13}$	0
MAE	$4.1 \times 10^{-10}$	$4.4 \times 10^{-16}$	$5.9 \times 10^{-13}$	$8.9 \times 10^{-16}$

Table 5.3: Comparison of the present method with Elbarbary's Chebyshev pseudospectral integration method (Elbarbary, 2007). The results are the observed MAEs at each collocation node  $x_i$ . The results of the present method are reported at  $\alpha = 0.6; 1$  for  $N = 8; 10$ , respectively.

**Example 5.3.5.** Consider the following linear fourth-order BVP:

$$y^{(4)} + xy = -(8 + 7x + x^3)e^x, \quad (5.67)$$

with the boundary conditions

$$y(0) = 0, y'(0) = 1, y''(1) = -4e, y'''(1) = -9e. \quad (5.68)$$

The exact solution is  $y(x) = x(1 - x)e^x$ .

Let  $r(x) = -(8 + 7x + x^3)e^x$ , then the HGIM transforms the problem into a linear system of equations of the form  $Aw = b$ , where the entries of the coefficient matrix  $A = (a_{ij})$  and the column vector  $b = (b_i)$  are given by

$$a_{ij} = \delta_{ij} + \frac{1}{6}x_j(6\hat{p}_{ij}^{(4)} - x_i^2((x_i - 3)\hat{p}_{N+1,j}^{(1)} + 3\hat{p}_{N+1,j}^{(2)})), \quad (5.69)$$

$$b_i = -\frac{1}{6}x_i^3\left(\sum_{j=0}^M p_{N+1,j}^{(1)} r_{N+1,j} + 9e\right) + \frac{1}{2}x_i^2\left(\sum_{j=0}^M p_{N+1,j}^{(1)} r_{N+1,j} - \sum_{j=0}^M p_{N+1,j}^{(2)} r_{N+1,j} + 5e\right) + \sum_{j=0}^M p_{ij}^{(4)} r_{ij} + x_i, \quad i, j = 0, \dots, N. \quad (5.70)$$

This problem was recently solved by Zahra (2011) using a spline method based on an exponential spline function. Comparisons between the present HGIM and Zahra's sixth-order spline method (Zahra, 2011) are shown in Table 5.4. The

## Chapter 5

latter reports the MAEs obtained in both methods for  $N = 8, 16, 32; M = N$ . The results confirm that the Gegenbauer polynomials achieve higher-order approximations than the standard Chebyshev and Legendre polynomials for many suitable values of  $\alpha$ . Figure 5.6 shows the numerical behavior of the HGIM. In particular, Figure 5.6(b) shows the MAEs of the present method at the same values of  $N$ . The figure suggests that higher precision approximations are expected via discretizations at the positive values of  $\alpha$  for small values of  $N$ , while Gegenbauer discretizations at the negative and some positive values of  $\alpha$  share excellent results for large values of  $N$ .

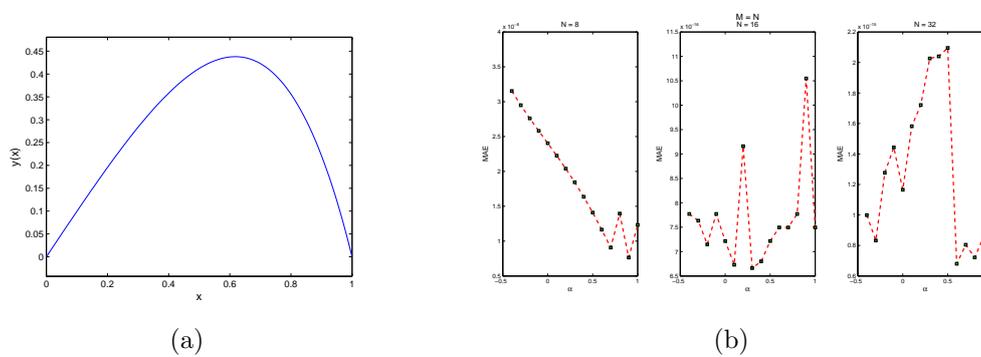


Figure 5.6: The numerical experiments of the HGIM on Example 5.3.5. Figure (a) shows the graph of  $y(x)$  on  $[0, 1]$ . Figure (b) shows the MAEs of the HGIM for  $N = 8, 16; 32$ .

Example 5.3.5		
$N$	Zahra's 6 <sup>th</sup> -order spline method (Zahra, 2011)	Present HGIM
8	$2.5316 \times 10^{-07}$	$7.6537 \times 10^{-09}$ (0.9)
16	$2.4800 \times 10^{-09}$	$6.6613 \times 10^{-16}$ (0.3)
32	$2.0891 \times 10^{-11}$	$6.8001 \times 10^{-16}$ (0.6)

Table 5.4: Comparison of the present method with Zahra's sixth-order spline method (Zahra, 2011).

**Example 5.3.6.** Consider the following Fredholm integral equation of the second kind:

$$y(x) - \int_{-1}^1 K(x,t)y(t)dt = f(x), \quad x \in [-1, 1], \quad (5.71)$$

## Chapter 5

---

with the kernel function  $K(x, t) = (x - t)^3 / (x^2(1 + t^2))$ ;

$$f(x) = \sqrt{1 + x^2} - \frac{3(\sqrt{2} - \arcsin h(1))}{x} - 2x \arcsin h(1).$$

The exact solution is  $y(x) = \sqrt{1 + x^2}$ .

Applying the HGIM leads to the following algebraic system of equations:

$$w_i - \sum_{j=0}^N \hat{p}_{N+1,j}^{(1)} K(x_i, x_j) w_j - f_i = 0, \quad i = 0, \dots, N, \quad (5.72)$$

where  $f_i = f(x_i) \forall i$ . Long et al. (2009) solved this problem with a multi-projection and iterated multi-projection methods using global polynomial bases— typically, Legendre polynomials were used as the orthonormal basis. The M-Galerkin and M-collocation methods employed lead to iterative solutions approximating the exact solution  $y$  with  $n^{-4k}$ -order of convergence in the supremum norm. Comparisons with Long et al. (2009) are shown in Table 5.5 for  $N = 3, 5, 7; 9$ . The results show that the present HGIM outperforms the M-Galerkin and M-collocation methods. Moreover, the present method enjoys the luxury of spectral convergence using a relatively small number of solution nodes. Figure 5.7(b) shows that high-precision approximations are obtained for Gegenbauer collocations at the negative values of  $\alpha$ , while degradation of precision can be observed for increasing values of  $\alpha$ . Furthermore, the figure manifests that the Gegenbauer polynomials generally perform better than the Chebyshev and Legendre polynomials. Nonetheless, it is interesting to note that the best approximations of the present method in the cases  $N = 5; 9$  are reported at  $\alpha = 0.5$  corresponding to the zeros of the Legendre polynomials. Also, the second better approximation for  $N = 7$  is reported at  $\alpha = 0.5$  (the best result is reported at  $\alpha = -0.4$ ). This suggests that Legendre polynomials can usually perform well for similar problems.

**Example 5.3.7.** Consider the following Fredholm integro-differential equation:

$$y'(x) - y(x) - \int_0^1 e^{sx} y(s) ds = \frac{1 - e^{x+1}}{x + 1}, \quad y(0) = 1, \quad (5.73)$$

with the exact solution  $y(x) = e^x$ .

Applying the HGIM results in the following algebraic system of linear equations:

$$w_i - \sum_{j=0}^N (\hat{p}_{ij}^{(1)} + \sum_{k=0}^N \hat{p}_{N+1,j}^{(1)} \hat{p}_{ik}^{(1)} e^{x_k x_j}) w_j - \sum_{j=0}^M p_{ij}^{(1)} r_{ij} - 1 = 0, \quad i = 0, \dots, N, \quad (5.74)$$

Example 5.3.6

$N$	Long et al.'s method (Long et al., 2009) M-Galerkin methods	Long et al.'s method (Long et al., 2009) M-collocation methods	Present HGIM
3	2.5798223000	$4.1346469 \times 10^{-01}$	$4.5289371 \times 10^{-04}$ (0.3)
5	$1.0458420 \times 10^{-01}$	$8.3487129 \times 10^{-03}$	$7.5585262 \times 10^{-05}$ (0.5)
7	$8.6205177 \times 10^{-04}$	$1.9766551 \times 10^{-04}$	$1.3571155 \times 10^{-06}$ (-0.4)
9	$8.7192025 \times 10^{-06}$	$9.0014073 \times 10^{-06}$	$8.7955126 \times 10^{-08}$ (0.5)

Table 5.5: Comparison of the present method with Long et al. (2009). The results are the observed MAEs of both methods.

## Chapter 5

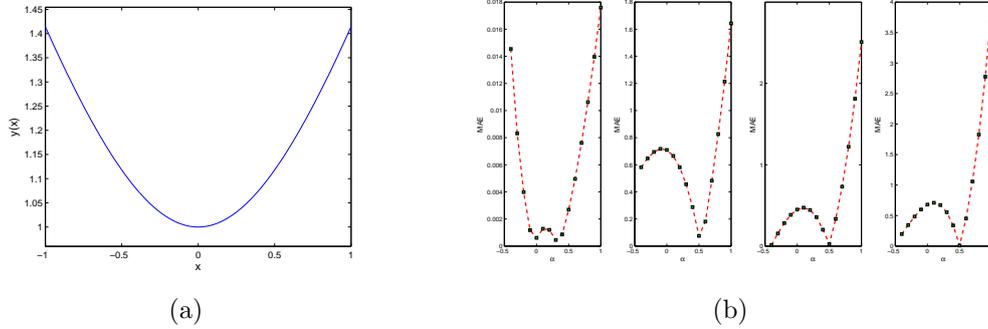


Figure 5.7: The numerical experiments of the HGIM on Example 5.3.6. Figure (a) shows the graph of  $y(x)$  on  $[-1, 1]$ . Figure (b) shows the MAEs of the HGIM for  $N = 3, 5, 7, 9$ .

where  $r(x) = (1 - e^{x+1})/(x + 1)$ . Figure 5.8(b) shows the MAEs of the HGIM for  $M = 14$ ;  $N = 3, 4, 6, 7, 9, 10$ . The best possible approximations obtained in the first five cases are reported at  $\alpha = 0.5$ , while the best approximation in the last case is reported at  $\alpha = 0.7$ . This problem was solved by Maleknejad and Attary (2011) using Shannon wavelets approximation based on Cattani's connection coefficients (Cattani, 2008). The Shannon wavelets expansions result in a linear system of dimension  $(N_1 + 2)(2M_1 + 1)$ , with  $(N_1 + 2)(2M_1 + 1)$  unknowns, where  $N_1, M_1$  are some parameters referring to the numbers of the terms in the Shannon scaling functions and mother wavelets expansions. Comparisons between the HGIM using  $M = 14$  and Maleknejad and Attary's method (Maleknejad and Attary, 2011) are shown in Table 5.6. The table demonstrates that the present method produces linear systems of lower dimensions than Maleknejad and Attary's method (Maleknejad and Attary, 2011), while achieving higher-order approximations. Hence the present HGIM may require less memory than alternative methods.

**Example 5.3.8.** Consider the following third-order integro-differential equation:

$$y'''(s) + \int_0^{\frac{\pi}{2}} s\tau y'(\tau) d\tau = \sin(s) - s, \quad (5.75)$$

with the initial conditions

$$y(0) = 1, y'(0) = 0, y''(0) = -1.$$

The exact solution is  $y(s) = \cos(s)$ .

Setting  $s = \pi(x + 1)/4, \tau = \pi(t + 1)/4$  transforms the problem into a third-order integro-differential equation defined on the domain  $[-1, 1]$ , and has the

## Chapter 5

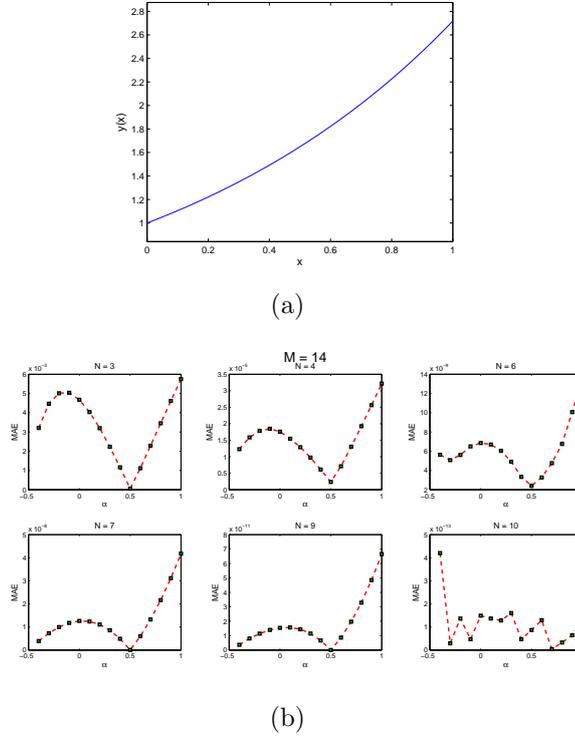


Figure 5.8: The numerical experiments of the HGIM on Example 5.3.7. Figure (a) shows the graph of  $y(x)$  on  $[0, 1]$ . Figure (b) shows the MAEs of the HGIM for  $M = 14$ ;  $N = 3, 4, 6, 7, 9, 10$ .

form

$$\left(\frac{4}{\pi}\right)^3 y'''(x) + \left(\frac{\pi}{4}\right)^2 \int_{-1}^1 (x+1)(t+1)y'(t)dt = \sin\left(\frac{\pi}{4}(x+1)\right) - \frac{\pi}{4}(x+1), \quad (5.76)$$

with the conditions

$$y(-1) = 1, y'(-1) = 0, 16y''(-1) = -\pi^2.$$

The exact solution becomes  $y(x) = \cos(\pi(x+1)/4)$ . The HGIM results in a  $(N+2) \times (N+2)$  algebraic linear system of the form  $Aw = b$ , where the entries of the coefficient matrix  $A = (a_{ij})$ , and the column vector  $b = (b_i)$  are given by

$$a_{ij} = \begin{cases} \left(\frac{4}{\pi}\right)^3 \delta_{ij} + \left(\frac{\pi}{4}\right)^2 \sum_{l=0}^M p_{il}^{(3)}(1+z_{il})(2\delta_{N+1,j} - \hat{p}_{N+1,j}^{(1)}), & 0 \leq j \leq N, \\ \left(\frac{4}{\pi}\right)^3 \delta_{ij} + \frac{\pi^2}{8} \sum_{l=0}^M p_{il}^{(3)}(1+z_{il})\delta_{N+1,j}, & j = N+1, \end{cases} \quad (5.77)$$

## Chapter 5

Example 5.3.7				
Maleknejad and Attary's method (Maleknejad and Attary, 2011)		MAE	Present HGIM	
$N_1$	$M_1$		$N$	MAE
2	2	$1.23 \times 10^{-04}$	3	$5.88 \times 10^{-05}$ (0.5)
2	3	$2.95 \times 10^{-06}$	4	$2.35 \times 10^{-06}$ (0.5)
4	3	$4.98 \times 10^{-09}$	6	$2.41 \times 10^{-09}$ (0.5)
5	2	$4.30 \times 10^{-10}$	7	$6.48 \times 10^{-11}$ (0.5)
8	3	$9.28 \times 10^{-13}$	9	$4.24 \times 10^{-14}$ (0.5)
9	3	$1.77 \times 10^{-14}$	10	$4.44 \times 10^{-15}$ (0.7)

Table 5.6: Comparison of the present method with Maleknejad and Attary's method (Maleknejad and Attary, 2011). The results are the observed MAEs in both methods.

$$b_i = \sum_{j=0}^M p_{ij}^{(3)} r_{ij} - \frac{2}{\pi}(1+x_i)^2 + \left(\frac{4}{\pi}\right)^3, \quad i = 0, \dots, N+1, \quad (5.78)$$

where  $r(x) = \sin(\pi(x+1)/4) - \pi(x+1)/4$ ;  $z_{il} \in S_{N+1,M}$  defined by Eq. (5.19). The P-matrix introduced recently by Elgindy and Smith-Miles (2013b) has been demonstrated to produce higher-order approximations by simply increasing the number of its columns without the need to increase the number of the solution nodes. This useful trick is illustrated in Figure 5.9(b), which shows the rapid convergence of the HGIM for  $M = 16$ ;  $N = 3, 7, 11, 15$ . This problem was solved by El-Kady et al. (2009) using a Gegenbauer spectral method. The method performs approximations to the highest order derivative in the linear integro-differential equations and generates approximations to the lower order derivatives through integration of the highest-order derivative. The resulting linear system is then modeled as a mathematical programming problem solved using the partial quadratic interpolation method (El-Gindy and Salim, 1990). Comparisons with El-Kady et al.'s Gegenbauer integration method (El-Kady et al., 2009) are shown in Table 5.7. The higher-order approximations obtained through the HGIM are clearly evident from the table even for a relatively small number of solution points. In fact, for a total number of solution points  $N_T = 8$ , for instance, the HGIM produces approximations of order  $O(16)$  twice the value of  $N_T$ , i.e. the approximations are accurate to almost full machine precision. Therefore the precision of the Gegenbauer polynomial approximations as clearly seen from the table can considerably exceed those obtained from both the Chebyshev and

## Chapter 5

Legendre polynomials. Moreover, the table shows for many different values of  $N$  that the Gegenbauer polynomials are very effective in the solution of high-order integro-differential equations.

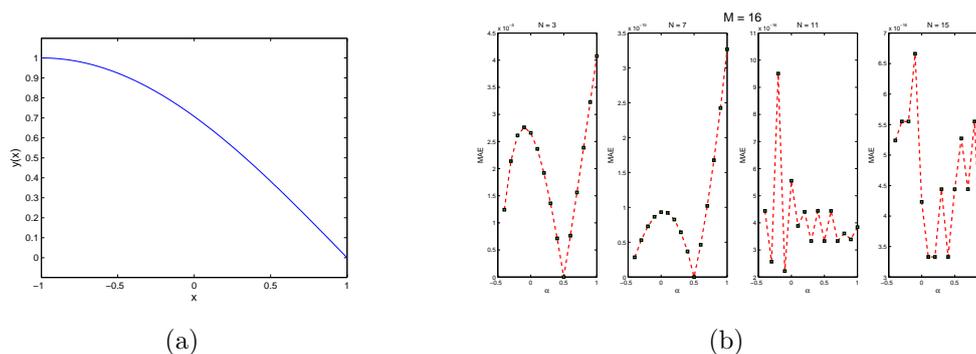


Figure 5.9: The numerical experiments of the HGIM on Example 5.3.8. Figure (a) shows the graph of  $y(x)$  on  $[-1, 1]$ . Figure (b) shows the MAE of the Gegenbauer spectral method for  $M = 16$ , and  $N = 3, 7, 11, 15$ .

Example 5.3.8		
$N_T$	El-Kady et al.'s Gegenbauer integration method (El-Kady et al., 2009)	Present HGIM
4	$3.10421 \times 10^{-04}$	$4.38620 \times 10^{-09}$ (0.5)
8	$4.23487 \times 10^{-09}$	$2.49800 \times 10^{-16}$ (0.5)
12	$1.06581 \times 10^{-14}$	$2.22045 \times 10^{-16}$ (-0.1)
16	$1.11022 \times 10^{-15}$	$3.33067 \times 10^{-16}$ (0.1, 0.2, 0.4; 0.9)

Table 5.7: Comparison of the present method with El-Kady et al.'s Gegenbauer integration method (El-Kady et al., 2009).  $N_T$  denotes the total number of nodes. The results are the observed MAEs in both methods.

## 5.4 Concluding Remarks

This chapter reports an efficient numerical method for solving BVPs, integral and integro-differential equations using GIMs. The key idea is to transform the general BVPs and integro-differential equations into their integral reformulations,

## Chapter 5

---

and then discretize using GIMs. The resulting algebraic linear system of equations can be solved for the solution values in the physical space using efficient linear system solvers. The proposed HGIM applies GIMs which generally lead to well-conditioned linear systems, and avoid the degradation of precision caused by severely ill-conditioned SDMs. The algorithm presented is numerically stable, and spectral accuracy is achieved using a relatively small number of solution points, which is a desired feature for a spectral method. The proposed HGIM has the ability to obtain higher-order approximations without the need to increase the number of the solution points. The applicability of the proposed method is illustrated via eight test examples. The obtained results are very consistent, with the performance of the proposed method superior to other competitive techniques in the recent literature regarding accuracy and convergence rate. Moreover, the developed Gegenbauer integration scheme is memory-minimizing and can be easily programmed. Furthermore, the chapter suggests that the Gegenbauer polynomials can generally perform better than their subclasses including Chebyshev and Legendre polynomials on a wide variety of problems. The present HGIM is broadly applicable and can be applied for solving many problems such as BVPs, integral and integro-differential equations, optimization problems, optimal control problems, etc.

This page is intentionally left blank

## PART B: Suggested Declaration for Thesis Chapter

Monash University

### Declaration for Thesis Chapter 6

Declaration by candidate

In the case of Chapter 6, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
The author of the key ideas, programming codes, organization, development, and writing up of the article	85

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided valuable comments and aided proofreading	
Boris Miller	Provided valuable comments and aided proofreading. Also suggested to study the UAV problem	

Candidate's Signature

	Date 11/05/2013
---	--------------------

### Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors;
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)	School of Mathematical Sciences, Monash University, Clayton Campus	
Signature 1		Date 14/5/13
Signature 2		13/05/2013

This page is intentionally left blank

## *Chapter 6*

# *Solving Optimal Control Problems Using a Gegenbauer Transcription Method*

**Chapter 6 is based on the article Elgindy, K. T., Smith-Miles, K. A., and Miller, B., 15–16 November 2012. Solving optimal control problems using a Gegenbauer transcription method. In: The Proceedings of 2012 Australian Control Conference, AUCC 2012. Engineers Australia, University of New South Wales, Sydney, Australia.**

**Abstract.** *In this chapter we describe a novel direct optimization method using Gegenbauer-Gauss (GG) collocation for solving continuous-time optimal control (OC) problems (CTOCPs) with nonlinear dynamics, state and control constraints. The time domain is mapped onto the interval  $[0, 1]$ , and the dynamical system formulated as a system of ordinary differential equations is transformed into its integral formulation through direct integration. The state and the control variables are fully parameterized using Gegenbauer expansion series with some unknown Gegenbauer spectral coefficients. The proposed Gegenbauer transcription method (GTM) then recasts the performance index, the reduced dynamical system, and the constraints into systems of algebraic equations using optimal Gegenbauer quadratures. Finally, the GTM transcribes the infinite-dimensional OC problem into a parameter nonlinear programming (NLP) problem which can be solved in the spectral space; thus approximating the state and the control variables along the entire time horizon. The high precision and the spectral convergence of the discrete solutions are verified through two OC test problems with nonlinear dynamics and some inequality constraints. The present GTM offers many useful properties and a viable alternative over the available direct optimization methods.*

References are considered at the end of the thesis.

# Chapter 6

## Solving Optimal Control Problems Using a Gegenbauer Transcription Method

### 6.1 Introduction

Optimal control (OC) theory is an elegant mathematical tool for making optimal decision policies pertained to complex dynamical systems. The theory plays an increasingly important role in the design and modelling of modern systems with broad attention from industry. The main goal of OC theory is to determine the input OC signals which influence a certain process to satisfy some physical constraints while optimizing some performance criterion. Although extensive studies have been conducted on OC problems governed by nonlinear dynamical systems, determining the OC within high-precision is still challenging for many problems. Classical solution methods such as the calculus of variations, dynamic programming and Pontryagin's maximum/minimum principle can provide the OC only in very special cases, but in general, a closed form expression of the OC is usually out of reach and not even practical to obtain (Bertsekas, 2005; Gong et al., 2006a). These difficulties drove the researchers and scholars, since last century, to search for viable alternative computational techniques to the aforementioned classical methods, taking advantage of the giant evolution in the areas of numerical analysis and approximation theory, and the advent of rapid and powerful digital computers. Among the available computational methods for solving continuous-time OC problems (CTOCPs), direct optimization methods convert the CTOCP into a finite-dimensional nonlinear programming (NLP) problem through control and/or state parameterization. The reduced NLP problem can be solved using the available robust optimization solvers. The simplicity of the discretization

## Chapter 6

---

procedure, the high accuracy, and the fast convergence of the solutions of the discretized OC problem have made the direct optimization methods the ideal methods of choice (Gong et al., 2008; Hesthaven et al., 2007), and well suited for solving OC problems, cf. (Betts, 2009; Elnagar and Kazemi, 1995; Fahroo and Ross, 2008; Garg et al., 2011a; Kang et al., 2008), and the references therein.

Our goal in this chapter is to develop an efficient direct optimization method for solving CTOCPs. Moreover, we aim to establish a high-order numerical scheme which results in a NLP problem with considerably low-dimension space to facilitate the application of the NLP solvers, and accelerate the solution procedure. The proposed method converts the CTOCP into a NLP problem through the application of a spectral collocation scheme based on Gegenbauer polynomials. To overcome the ill-conditioning of the spectral differentiation matrices (SDMs), the underlying dynamical system of the differential equations is transformed into its integral formulation through direct integration. Both the state and the control variables are parameterized and approximated by truncated Gegenbauer expansion series with unknown Gegenbauer spectral coefficients. The time domain is discretized at the Gegenbauer-Gauss (GG) points. The integral operations are approximated using the optimal Gegenbauer operational matrices of integration known as the optimal P-matrices (see the appendix). The developed technique reduces the cost function, the dynamics, and the constraints into systems of algebraic equations; thus greatly simplifying the problem. In this manner, the infinite-dimensional OC problem is transcribed into a finite-dimensional NLP problem, which can be solved in the Gegenbauer spectral space using the well-developed NLP techniques and computer codes. The remaining of the chapter is organized as follows: In Section 6.2 we present the CTOCP statement. In Section 6.3 we introduce the Gegenbauer transcription method (GTM) for the solution of CTOCPs. Two numerical experiments are presented in Section 6.4 to demonstrate the efficiency and the spectral accuracy of the proposed method followed by a discussion and some concluding remarks in Section 6.5. A brief background on the Gegenbauer polynomials and their associated quadratures is provided in the appendix.

### 6.2 The OC Problem Statement

Consider, without loss of generality, the following nonlinear CTOCP ( $P_1$ ) with fixed final time, mixed state-control path and terminal inequality constraints:

$$\text{minimize } J(u(t)) = \Phi(x(1)) + \int_0^1 \mathcal{L}(x(\tau), u(\tau), \tau) d\tau, \quad (6.1a)$$

$$\text{subject to } \dot{x}(t) = f(x(t), u(t), t), \quad (6.1b)$$

$$x(0) = x^0, \quad (6.1c)$$

$$\psi_i(x(t), u(t), t) \leq 0, \quad i = 0, \dots, \ell; \quad (6.1d)$$

$$\phi(x(1), u(1)) \leq 0. \quad (6.1e)$$

Problem ( $P_1$ ) is known as the Bolza problem, where  $[0, 1]$  is the time interval of interest,  $x : [0, 1] \rightarrow \mathbb{R}^n$  is the state vector,  $\dot{x} : [0, 1] \rightarrow \mathbb{R}^n$  is the vector of first-order time derivatives of the states,  $u : [0, 1] \rightarrow \mathbb{R}^m$  is the control vector,  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$  is the terminal cost function,  $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}$  is the Lagrangian function,  $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^n$  is a vector field, where each system function  $f_i$  is continuously differentiable with respect to  $x$ , and is continuous with respect to  $u$ . Both functions  $\Phi$  and  $\mathcal{L}$  are continuously differentiable with respect to  $x$ ;  $\mathcal{L}$  is continuous with respect to  $u$ .  $J$  is the cost function to be minimized. Equations (6.1b) & (6.1c) represent the dynamics of the system and its initial state condition.  $\psi_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}$  is an inequality constraint on the state and the control vectors for each  $i$ .  $\phi : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$  is a terminal inequality constraint on the state and the control vectors. We shall assume that for any admissible control trajectory  $u(t)$ , the dynamical system has a unique state trajectory  $x(t)$ . The goal of Problem ( $P_1$ ) is to determine the optimal admissible control policy  $u(t)$  in the time horizon  $[0, 1]$  such that  $J$  is minimized. In general, a CTOCP defined over the physical time domain  $[t_0, t_f]$  can be reformulated into the Bolza Problem ( $P_1$ ) using the strict change of variable  $t = (\tau - t_0)/(t_f - t_0)$ , where  $\tau \in [t_0, t_f]$ ;  $t_0$  and  $t_f$  are the initial and final times, respectively.

### 6.3 The GTM

In this section we shall describe a novel method for the numerical solution of nonlinear CTOCPs formulated in the preceding section based on GG collocation. The numerical scheme involves the approximation of the system dynamics, where it is necessary to approximate the derivatives of the state variables at the GG points. This can be accomplished through SDMs. Nevertheless, SDMs are known to be severely ill-conditioned (Elbarbary, 2006, 2007; Funaro, 1987), and their implementation causes degradation of the observed precision (Tang and Trummer,

## Chapter 6

---

1996). Moreover, it has been shown that the time step restrictions can be more severe than those predicted by the standard stability theory (Trefethen, 1988; Trefethen and Trummer, 1987). For higher-order SDMs, the ill-conditioning becomes very critical to the extent that developing efficient preconditioners is extremely crucial (Elbarbary, 2006; Hesthaven, 1998). Another approach is to transform the dynamical system into its integral formulation, where the state and the control variables are approximated by truncated spectral expansion series, while the integral operations are approximated by spectral integration matrices (SIMs). This numerical scheme is generally well-behaved and associated with many advantages: (i) SIMs are known to be well-conditioned operators (Elbarbary, 2006, 2007; Elgindy, 2009; Elgindy and Smith-Miles, 2013b,c; Greengard, 1991; Lundbladh et al., 1992). (ii) The well-conditioning of SIMs is essentially unaffected for increasing number of grid points (Elgindy, 2009; Elgindy and Smith-Miles, 2013c). (iii) Greengard and Rokhlin (1991) showed that the integral equation formulation, when applied to two-point boundary value problems (TPBVPs) for instance, is insensitive to boundary layers, insensitive to end-point singularities, and leads to small condition numbers while achieving high computational efficiency. (iv) The use of integration for constructing the spectral approximations improves the rate of convergence of the spectral interpolants, and allows the multiple boundary conditions to be incorporated more efficiently (Elgindy, 2009; Elgindy and Smith-Miles, 2013b,c; Mai-Duy and Tanner, 2007). These useful features in addition to the promising results obtained by Elgindy and Smith-Miles (2013b,c) motivate us to apply the Gegenbauer integration scheme for approximating the underlying dynamical system of the CTOCP.

Integrating Equation (6.1b) and using the initial condition (6.1c) recasts the dynamical system into its integral formulation given by

$$x(t) = \int_0^t f(x(\tau), u(\tau), \tau) d\tau + x^0. \quad (6.2)$$

Equation (6.2) together with Equations (6.1a), (6.1d); (6.1e) represent the integral Bolza problem (P<sub>2</sub>). To approximate the CTOCP, we seek Gegenbauer polynomial expansions of the state and the control variables in the form

$$x_r(t) \approx \sum_{k=0}^L a_{rk} C_k^{(\alpha)}(t), \quad r = 1, \dots, n, \quad (6.3a)$$

$$u_s(t) \approx \sum_{k=0}^M b_{sk} C_k^{(\alpha)}(t), \quad s = 1, \dots, m, \quad (6.3b)$$

and collocate at the GG points  $t_i \in S_N^{(\alpha)} = \{t_i | C_{N+1}^{(\alpha)}(t_i) = 0, i = 0, \dots, N\}$ . Let  $t_{N+1} = 1$ ,  $\hat{e}_l \in \mathbb{R}^{l+1} : (\hat{e}_l)_k = 1$ ,  $z_{i'j} \in S_{N+1, M_P}$ ,  $a = (a_1, \dots, a_n)^T$ ,  $b =$

## Chapter 6

---

$(b_1, \dots, b_m)^T, a_r = (a_{r0}, \dots, a_{rL})^T, b_s = (b_{s0}, \dots, b_{sM})^T, \xi_l^{(\alpha)} \in \mathbb{R}^{(N+1) \times (l+1)} :$   
 $(\xi_l^{(\alpha)})_{ik} = (C_k^{(\alpha)}(t_i)), \hat{\xi}_{li}^{(\alpha)T} \in \mathbb{R}^{l+1} : (\hat{\xi}_{li}^{(\alpha)})_k = (\xi_l^{(\alpha)})_{ik}, \zeta_l^{(\alpha)} \in \mathbb{R}^{(N+2) \times (M_P+1) \times (l+1)} :$   
 $(\zeta_l^{(\alpha)})_{i'jk} = (C_k^{(\alpha)}(z_{i'j})), \hat{\zeta}_{li'j}^{(\alpha)T} \in \mathbb{R}^{l+1} : (\hat{\zeta}_{li'j}^{(\alpha)})_k = (\zeta_l^{(\alpha)})_{i'jk}, r = 1, \dots, n; s =$   
 $1, \dots, m; i = 0, \dots, N; i' = 0, \dots, N+1; j = 0, \dots, M_P; k = 0, \dots, l; M_P, l \in \mathbb{Z}^+.$   
Hence the state and the control vectors at the GG solution points can be written as

$$x(t_i) \approx (I_n \otimes \hat{\xi}_{Li}^{(\alpha)})a, \quad (6.4a)$$

$$u(t_i) \approx (I_m \otimes \hat{\xi}_{Mi}^{(\alpha)})b, \quad (6.4b)$$

where  $I_l$  is the identity matrix of order  $l$ ;  $\otimes$  is the Kronecker product of matrices. Using Equation (6.21), we can show that  $x(1) = (I_n \otimes \hat{e}_L^T)a; u(1) = (I_m \otimes \hat{e}_M^T)b$ . Hence the discrete approximation of the cost function can be represented by

$$J \approx \tilde{J}(a, b) = \Phi((I_n \otimes \hat{e}_L^T)a) + P_{N+1}^{(1)}\hat{\mathcal{L}}, \quad (6.5)$$

where  $\hat{\mathcal{L}} \in \mathbb{R}^{M_P+1} : (\hat{\mathcal{L}})_j = \mathcal{L}((I_n \otimes \hat{\zeta}_{L, N+1, j}^{(\alpha)})a, (I_m \otimes \hat{\zeta}_{M, N+1, j}^{(\alpha)})b, z_{N+1, j}); j = 0, \dots, M_P$ . The discrete dynamical system at the GG grid points is given by

$$\mathcal{H}_i(a, b) = (I_n \otimes \hat{\xi}_{Li}^{(\alpha)})a - (I_n \otimes P_i^{(1)})F_i - x^0 \approx 0, \quad i = 0, \dots, N, \quad (6.6)$$

where  $F_i = (F_{1i}, \dots, F_{ni})^T, F_{ri} = (f_{ri0}, \dots, f_{riM_P})^T, f_{rij} = f_r((I_n \otimes \hat{\zeta}_{Lij}^{(\alpha)})a, (I_m \otimes \hat{\zeta}_{Mij}^{(\alpha)})b, z_{ij}), r = 1, \dots, n; i = 0, \dots, N; j = 0, \dots, M_P$ . The discrete approximations of the inequality constraints (6.1d) and the terminal constraint (6.1e) become

$$c_{ij}(a, b) = \psi_j((I_n \otimes \hat{\xi}_{Li}^{(\alpha)})a, (I_m \otimes \hat{\xi}_{Mi}^{(\alpha)})b, t_i) \leq 0, \quad i = 0, \dots, N; j = 0, \dots, \ell; \quad (6.7)$$

$$c_t(a, b) = \phi((I_n \otimes \hat{e}_L^T)a, (I_m \otimes \hat{e}_M^T)b) \leq 0, \quad (6.8)$$

respectively. Hence the CTOCP (P<sub>1</sub>) is reduced to the following parameter NLP (P<sub>3</sub>):

$$\text{minimize } \tilde{J}(a, b), \quad (6.9a)$$

$$\text{subject to } \mathcal{H}_i(a, b) = 0, \quad (6.9b)$$

$$c_{ij}(a, b) \leq 0, \quad i = 0, \dots, N; j = 0, \dots, \ell; \quad (6.9c)$$

$$c_t(a, b) \leq 0. \quad (6.9d)$$

Problem (P<sub>3</sub>) can be solved using well-developed optimization software for the Gegenbauer coefficients  $a; b$ . The state and the control variables can then be evaluated at the GG nodes using Equations (6.4a) & (6.4b). In fact, since the GTM

## Chapter 6

---

solves the CTOCP ( $P_1$ ) in the spectral space, the approximation can immediately be evaluated at any time history of both the control and the state variables once the approximate spectral coefficients are found without invoking any interpolation method. This is a clear advantage over the “classical” discretization methods such as the finite difference schemes, which require a further step of interpolation to evaluate an approximation at an intermediate point. This useful feature establishes the power of the proposed GTM for solving CTOCPs as the optimal state and control profiles are readily determined.

### 6.4 Illustrative Numerical Examples

**Example 6.4.1** (The path planning in a threat environment). Consider the problem of determining the OC  $\gamma(t) \in C[0, T]$  which minimizes the performance index

$$J = \int_0^T f(x(t), y(t)) dt, \quad (6.10a)$$

$$\text{subject to } \dot{x}(t) = V \cos(\gamma(t)), \quad (6.10b)$$

$$\dot{y}(t) = V \sin(\gamma(t)), \quad (6.10c)$$

$$|\gamma(t)| \leq \pi, \quad (6.10d)$$

$$V \in [V_1, V_2], \quad (6.10e)$$

$$T \in [T_1, T_2], \quad (6.10f)$$

$$(x(0), y(0)) = (-3, -4), \quad (6.10g)$$

$$(x(T), y(T)) = (3, 3), \quad (6.10h)$$

where

$$\begin{aligned} f(x, y) = & 4/((x + 1.3)^2 + (y + 1.3)^2 + 1) + 2/((x - 1.9)^2 + (y - 1.6)^2 + 1) \\ & + 1/((x - 0.4)^2 + (y - 0.1)^2 + 1) + 1/((x - 0.6)^2 + (y - 0.1)^2 + 0.5), \end{aligned} \quad (6.10i)$$

$V_1, V_2, T_1, T_2$  are some real parameters such that  $V_1 < V_2; T_1 < T_2$ .

Example 6.4.1 was studied by Miller et al. (2011), and serves as a model case for a CTOCP governed by a nonlinear dynamical system with boundary conditions and control constraints. The OC model asks for the optimal path planning in 2D for an unmanned aerial vehicle (UAV) mobilizing in a stationary risk environment. The control  $\gamma(t)$  is the yaw angle constrained by Inequality (6.10d). The running cost  $f$  represents the hazard rate along the path (the stationary threats relief).  $T$  is a free variable and represents the flight time;  $V$

## Chapter 6

---

is a constant linear velocity satisfying Equation (6.10e).  $A = (-3, -4); B = (3, 3)$  are the initial and terminal conditions, respectively. The nonlinearity of the dynamical system, the boundary conditions and the control path constraint represent a very challenging task and add more complexity both analytically and computationally. Therefore the implementation of efficient and advanced numerical discretization schemes is of utmost importance.

Through the change of variable  $s = t/T$ , the CTOCP can be restated as follows:

$$\text{minimize } J = T \int_0^1 f(\tilde{x}(s), \tilde{y}(s)) ds, \quad (6.11a)$$

$$\text{subject to } \dot{\tilde{x}}(s) = TV \cos(\gamma(sT)) = \tilde{V} \cos(\tilde{\gamma}(s)), \quad (6.11b)$$

$$\dot{\tilde{y}}(s) = TV \sin(\gamma(sT)) = \tilde{V} \sin(\tilde{\gamma}(s)), \quad (6.11c)$$

$$|\tilde{\gamma}(s)| \leq \pi, \quad (6.11d)$$

$$\tilde{V} \in [V_1 T, V_2 T], \quad (6.11e)$$

$$(\tilde{x}(0), \tilde{y}(0)) = (-3, -4), \quad (6.11f)$$

$$(\tilde{x}(1), \tilde{y}(1)) = (3, 3), \quad (6.11g)$$

where  $\tilde{x}(s) = x(sT); \tilde{y}(s) = y(sT)$ . Hence if we consider the auxiliary cost function

$$J^{\text{aux}} = J/T = \int_0^1 f(\tilde{x}(s), \tilde{y}(s)) ds, \quad (6.12)$$

and the dynamical system equations (6.11b) and (6.11c) together with Conditions (6.11d)-(6.11g), then the cost function  $J$  and the auxiliary cost function  $J^{\text{aux}}$  are related by (Miller et al., 2011)

$$\min_{V, T, \gamma} J(V, T, \gamma) = \min_V \left( \frac{1}{V} \min_{\tilde{V}} \min_{\tilde{\gamma}} J^{\text{aux}}(\tilde{V}, \tilde{\gamma}) \right). \quad (6.13)$$

Equations (6.12), (6.11b)-(6.11g) represent the auxiliary 2D path planning problem (P<sub>4</sub>). Notice here that the velocity  $V$  may be viewed as an added control parameter influencing the values of the angular velocities in reasonable limits. The best path corresponding to the minimum risk value is the path corresponding to the minimum value of the product  $\tilde{V} J^{\text{aux}}$ . The optimal paths of the original problem are then obtained from the relations  $x(t) = \tilde{x}(t/T), y(t) = \tilde{y}(t/T), \gamma(t) = \tilde{\gamma}(t/T)$ , with  $T = \tilde{V}/V; \min J = T \min J^{\text{aux}}$ . Using a penalization approach, Miller et al. (2011) were able to recast the CTOCP into a TPBVP, which was solved numerically using MATLAB software. The reported risk integral value was found to be  $J = T J^{\text{aux}} \approx 3.1$ .

The implementation of the GTM presented in Section 6.3 involves three main stages:

## Chapter 6

---

(i) Recasting the dynamical system into its integral formulation given by

$$\tilde{x}(s) = \tilde{V} \int_0^s \cos(\tilde{\gamma}(\tau)) d\tau - 3; \quad (6.14a)$$

$$\tilde{y}(s) = \tilde{V} \int_0^s \sin(\tilde{\gamma}(\tau)) d\tau - 4. \quad (6.14b)$$

(ii) The approximation of the state and the control variables by the Gegenbauer expansion series

$$\tilde{x}(s) \approx \sum_{k=0}^L a_{1,k} C_k^{(\alpha)}(s), \quad (6.15a)$$

$$\tilde{y}(s) \approx \sum_{k=0}^L a_{2,k} C_k^{(\alpha)}(s); \quad (6.15b)$$

$$\tilde{\gamma}(s) \approx \sum_{k=0}^M b_k C_k^{(\alpha)}(s). \quad (6.15c)$$

(iii) The discretization of the time domain at the GG points  $s_i \in S_N^{(\alpha)}$ .

The GTM then transcribes the CTOCP (P<sub>4</sub>) into the following parameter finite-dimensional NLP problem (P<sub>5</sub>):

$$\text{minimize } J^{\text{aux}} \approx P_{N+1}^{(1)} \hat{f}, \quad (6.16a)$$

$$\text{subject to } (I_2 \otimes \hat{\xi}_{Li}^{(\alpha)}) a = \tilde{V} (I_2 \otimes P_i^{(1)}) \mathfrak{S}_i - (3, 4)^T, \quad (6.16b)$$

$$(I_2 \otimes \hat{e}_L^T) a = 3\hat{e}_1, \quad (6.16c)$$

$$\tilde{V} (I_2 \otimes P_{N+1}^{(1)}) \mathfrak{S}_{N+1} = (6, 7)^T, \quad (6.16d)$$

$$\left| \hat{\xi}_{Mi}^{(\alpha)} b \right| \leq \pi, \quad (6.16e)$$

where  $\mathfrak{S}_{i'} = (\cos(\chi_{Mi'}^{(\alpha)} b), \sin(\chi_{Mi'}^{(\alpha)} b))^T$ ,  $\chi_{Mi'}^{(\alpha)} \in \mathbb{R}^{(M_P+1) \times (M+1)} : (\chi_{Mi'}^{(\alpha)})_{jk} = (\zeta_M^{(\alpha)})_{i'jk}$ ,  $(\hat{f})_j = f(\hat{\zeta}_{L,N+1,j}^{(\alpha)} a_1, \hat{\zeta}_{L,N+1,j}^{(\alpha)} a_2)$ ,  $i' = 0, \dots, N+1$ ;  $i = 0, \dots, N$ ;  $j = 0, \dots, M_P$ ;  $k = 0, \dots, M$ . The reduced NLP (P<sub>5</sub>) can be solved for the Gegenbauer coefficients  $a_1 = (a_{1,0}, \dots, a_{1,L})^T$ ,  $a_2 = (a_{2,0}, \dots, a_{2,L})^T$ ;  $b = (b_0, \dots, b_M)^T$  using the available powerful optimization methods. Our numerical results for  $V = 2$  are shown in Table 6.1, where all calculations were performed on a personal laptop with a 2.53 GHz Intel Core i5 CPU and 4G memory running MATLAB 7.10.0.499 (R2010a) software in double precision real arithmetic. The numerical experiments were

## Chapter 6

implemented using MATLAB “fmincon” interior-point algorithm optimization solver with the solutions termination tolerance “TolX” =  $10^{-15}$ . The P-matrix was constructed via Algorithm 2.2 given in (Elgindy and Smith-Miles, 2013b) with  $M_P = M_{\max} = 20$ . The average CPU time in 100 runs taken by the GTM using  $\alpha = 0.2; N = L = M = 8$  was found to be 2.09 seconds. The  $(MAE)_{BC}$  was found to be  $7.1054 \times 10^{-15}$ , and the risk integral value  $J \approx 2.96$ . The risk integral value  $J$  then decreases for increasing values of the parameters  $N, L, M$ , and some suitable values of  $\alpha$ . For  $\alpha = -0.4, N = 9; L = M = 12$ , the calculated risk integral value was found to be  $J \approx 2.89$ . We notice that the approximate cost function values  $J$  obtained by the GTM are lower than the approximate value obtained in (Miller et al., 2011) for several values of  $\alpha, N, L; M$  with very accurate constraints satisfaction. This shows the greater accuracy and efficiency of the proposed GTM developed in our research work. The OC plot is shown in Figure 6.1(a), and the optimal state trajectory is shown in Figure 6.1(b) and Figure 6.2. The latter manifests that the best path for the UAV through the risk environment is to move around the threats’ centers and mountain terrain.

<b>Example 6.4.1</b>						
<b>Present GTM</b>						
	$\tilde{V}$	$J^{\text{aux}}$	$\tilde{V} J^{\text{aux}}$	$(MAE)_{BC}$	$T$	$J = T J^{\text{aux}}$
$L = M = 8$ at $S_8^{(0.2)}$	14.3768	0.4115	5.9164	$7.1054 \times 10^{-15}$	7.1884	2.9582
$L = M = 9$ at $S_9^{(-0.2)}$	14.3768	0.4042	5.8108	$4.4409 \times 10^{-16}$	7.1884	2.9054
$L = M = 10$ at $S_8^{(-0.1)}$	14.5610	0.3972	5.7839	$4.4409 \times 10^{-16}$	7.2805	2.8919
$L = M = 12$ at $S_9^{(-0.4)}$	14.3768	0.4017	5.7750	$8.88178 \times 10^{-16}$	7.1884	2.8875

Table 6.1: The results of the present GTM for  $V = 2$ , and different values of  $\alpha, N, L; M$ .  $(MAE)_{BC}$  denotes the MAE at the boundary condition (6.10h).

**Example 6.4.2.** Consider Example 6.4.1 with the following hazard relief function:

$$f(x, y) = 1/(x^2 + y^2 + 0.5) + 1/(x^2 + (y - 4)^2 + 1) + 2/((x - 2)^2 + (y + 4)^2 + 1). \quad (6.17)$$

This example is similar to Example 6.4.1 with a change in the mountains terrain localization and their slopes, where the new threats’ centers are  $(0, 0)^T, (0, 4)^T; (2, -4)^T$ . The application of the GTM leads to the reduced NLP (P<sub>5</sub>). Table 6.2 shows the GTM results for different values of  $\alpha, N, L; M$ . We notice here that the best results are reported at the GG discretization points set  $S_N^{(\alpha)}$  for  $\alpha < 0$ .

## Chapter 6

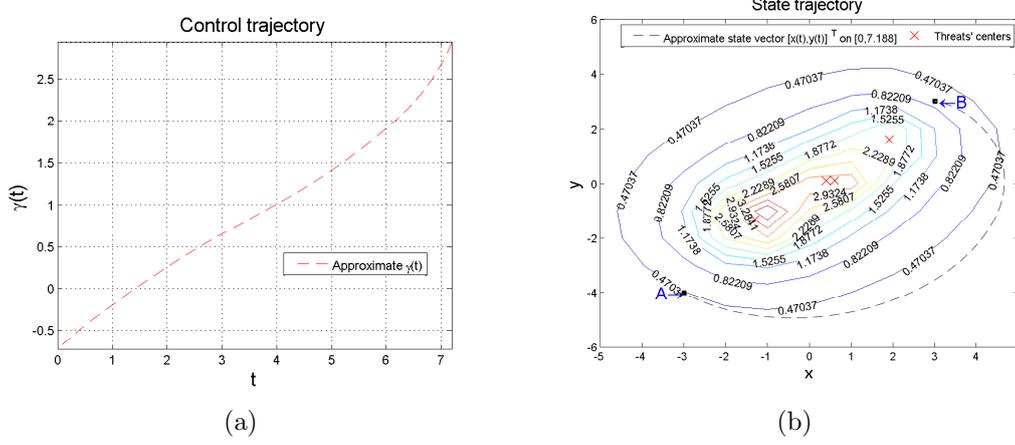


Figure 6.1: The numerical experiments of the GTM on Example 6.4.1. Figure (a) shows the profile of the control history on the calculated flight time domain  $[0, 7.188]$ . Figure (b) shows the 2D state trajectory in the hazard relief contour. The results are obtained at  $S_8^{(0.2)}$  using  $L = M = 8$ .

The average CPU time in 100 runs using  $\alpha = -0.1$ ;  $N = L = M = 5$  was found to be 0.7956 seconds. The  $(MAE)_{BC}$  was found to be  $4.885 \times 10^{-15}$ , and the risk integral value  $J \approx 1.44$ . Hence the adapted GTM for the solution of intricate CTOCPs reduces the required calculation time significantly while achieving very precise constraints satisfaction. The convergence of the GTM increases rapidly to the risk integral value  $J \approx 1.42$  for increasing number of the collocation points and the Gegenbauer expansion terms. The OC plot is shown in Figure 6.3(a), and the optimal state trajectory is shown in Figure 6.3(b) and Figure 6.4. In this example, the best path for the UAV through the risk environment is to move across the hazard mountains.

### Example 6.4.2

#### Present GTM

	$\tilde{V}$	$J^{\text{aux}}$	$\tilde{V} J^{\text{aux}}$	$(MAE)_{BC}$	$T$	$J = T J^{\text{aux}}$
$L = M = 5$ at $S_5^{(-0.1)}$	12.5351	0.2296	2.8782	$4.8850 \times 10^{-15}$	6.2676	1.4391
$L = M = 6$ at $S_6^{(-0.4)}$	13.2718	0.2146	2.8482	$2.0872 \times 10^{-14}$	6.6359	1.4241
$L = M = 8$ at $S_8^{(-0.3)}$	12.7193	0.2239	2.8476	$4.4409 \times 10^{-16}$	6.3596	1.4238
$L = M = 10$ at $S_{10}^{(-0.4)}$	11.9826	0.2375	2.8455	$8.8818 \times 10^{-16}$	5.9913	1.4227

Table 6.2: The results of the present GTM for  $V = 2$ , and different values of  $\alpha, N, L; M$ .

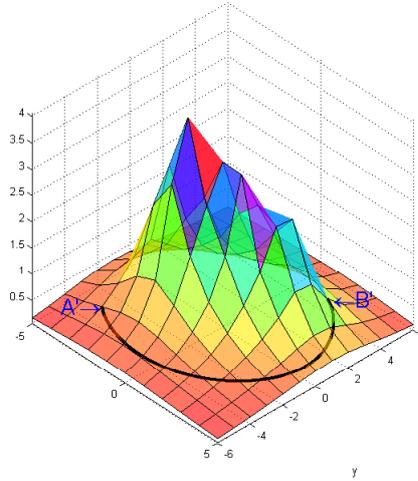
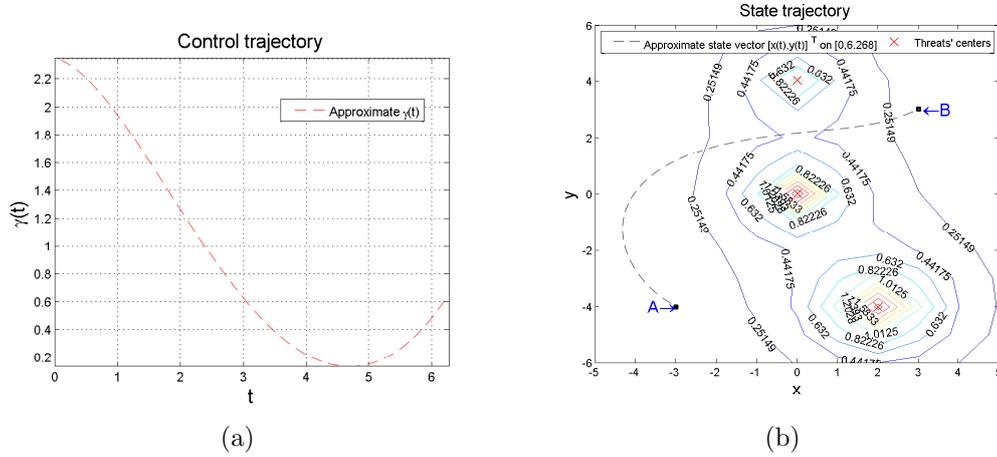


Figure 6.2: The figure shows the projected state trajectory in a black solid line along the 3D hazard relief.  $A'$ ;  $B'$  denote the points  $(-3, -4, f(-3, -4))$ ;  $(3, 3, f(3, 3))$ , respectively. The results are obtained at  $S_8^{(0.2)}$  using  $L = M = 8$ .

## 6.5 Discussion and Conclusion

The implementation of the GTM reveals many fruitful outcomes over the standard variational methods and direct collocation methods. The proposed GTM neither requires the explicit derivation and construction of the necessary conditions nor the calculation of the gradients  $\nabla_x \mathcal{L}$  of the Lagrangian function  $\mathcal{L}(x(t), u(t), t)$  w.r.t. the state variables, yet it is able to produce rapid convergence and achieve high precision approximations. In contrast, the indirect method applied by Miller et al. (2011) requires the explicit derivation of the adjoint equations, the control equations, and all of the transversality conditions. Moreover, the user must calculate the gradients  $\nabla_x \mathcal{L}$  for the solution of the necessary conditions of optimality. This useful feature of the GTM inherited from the direct optimization methods represents a significant contribution over the standard variational methods. From another standpoint, since the optimal P-matrix is constant for a particular GG solution points set, the GTM can be quickly used to solve many practical trajectory optimization problems. Moreover, decreasing the values of the parameters  $M_P$ ;  $M_{\max}$  required for the construction of the P-matrix via Algorithm 2.2 given in (Elgindy and Smith-Miles, 2013b) can reduce the calculations time taken by the GTM for solving CTOCPs with a slight reduction in accuracy. For instance, in Example 6.4.1, the GTM implemented using  $M_P = M_{\max} = 14$ , for  $\alpha = -0.2$ ;  $N = L = M = 8$ , produces the risk integral value  $J \approx 3$ , which is close to the value of  $J$  obtained in the first row of Table 6.1. The recorded average

## Chapter 6



## Chapter 6

---

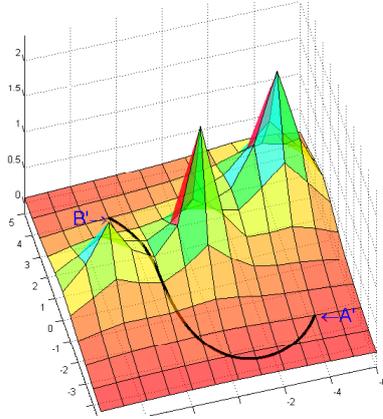


Figure 6.4: The figure shows the projected state trajectory in a black solid line along the 3D hazard relief.  $A'$ ;  $B'$  denote the points  $(-3, -4, f(-3, -4))$ ;  $(3, 3, f(3, 3))$ , respectively. The results are obtained at  $S_5^{(-0.1)}$  using  $L = M = 5$ .

the GTM cannot be realized in the presence of discontinuities and/or nonsmoothness in the OC problem, or in its solutions, as the convergence rate grows slower in this case for increasing number of the GG collocation points and the Gegenbauer expansion terms. Some research studies in this area manifest that the accuracies of direct global collocation methods and direct local collocation methods become comparable for nonsmooth OC problems, cf. (Huntington, 2007). To recover the exponential convergence property of the GTM in the latter case, the GTM can be applied within the framework of a semi-global approach. Here the OC problems can be divided into multiple-phases, which can be linked together via continuity conditions (linkage constraints) on the independent variable, the state, and the control. The GTM can then be applied globally within each phase. The reader may consult Ref. (Rao, 2003), for instance, for a similar practical implementation of this solution method. Another possible approach to accelerate the convergence rate of the GTM, and to recover the spectral accuracy, is to treat the GTM with an appropriate smoothing filter, cf. (Elnagar and Kazemi, 1998b), for instance, for a parallel approach using a pseudospectral Legendre method. Other methods include the knotting techniques developed in (Ross and Fahroo, 2002, 2004) for solving nonsmooth OC problems, where the dynamics are governed by controlled differential inclusions.

The work introduced in this chapter represents a major advancement in the area of direct orthogonal collocation methods using Gegenbauer polynomials. The simplicity and efficiency of the GTM allow for the implementation of a rapid and accurate trajectory optimization. The developed GTM can be easily extended

## Chapter 6

---

to higher dimensional OC problems under the same level of complexity, whereas the application of the variational methods leads to a TPBVP, which is rather unstable and hard to solve. Similar ideas to the one described in this chapter can be applied on CTOCPs governed by integral equations or integro-differential equations; therefore, the GTM encompasses a wider range of OC problems over the standard direct optimization methods.

In the numerical examples presented in this chapter, the results clearly show that the Gegenbauer polynomials are very effective in direct optimization techniques for solving hard OC problems. The reported results seem to favour the discretization of CTOCPs at the GG points for negative values of the Gegenbauer parameter  $\alpha$ ; however, better approximations using the GTM are possible for different choices of the GG discretization points. Further tests and analysis are necessary to investigate the stability, the accuracy, and the convergence of the method to the solution of CTOCPs. Finally, the present GTM offers many useful properties, and provides a strong addition to the arsenal of direct optimization methods.

## APPENDIX

### 6.5.1 Elementary Properties and Definitions

The Gegenbauer polynomial  $C_n^{(\alpha)}(x)$ ,  $n \in \mathbb{Z}^+$  of degree  $n$  and associated with the real parameter  $\alpha > -1/2$  is a real-valued function, which appears as an eigensolution to the singular Sturm-Liouville problem in the finite domain  $[-1, 1]$  (Szegö, 1975):

$$\frac{d}{dx}(1-x^2)^{\alpha+\frac{1}{2}} \frac{dC_n^{(\alpha)}(x)}{dx} + n(n+2\alpha)(1-x^2)^{\alpha-\frac{1}{2}} C_n^{(\alpha)}(x) = 0. \quad (6.18)$$

The weight function for the Gegenbauer polynomials is the even function  $(1-x^2)^{\alpha-1/2}$ . The form of the Gegenbauer polynomials is not unique, and depends on a certain standardization. The Gegenbauer polynomials standardized by Doha (1990) so that

$$C_n^{(\alpha)}(x) = \frac{n! \Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} P_n^{(\alpha-\frac{1}{2}, \alpha-\frac{1}{2})}(x), \quad n = 0, 1, 2, \dots, \quad (6.19)$$

establish the following useful relations:  $C_n^{(0)}(x) = T_n(x)$ ,  $C_n^{(1/2)}(x) = L_n(x)$ ;  $C_n^{(1)}(x) = (1/(n+1))U_n(x)$ , where  $P_n^{(\alpha-\frac{1}{2}, \alpha-\frac{1}{2})}(x)$  is the Jacobi polynomial of degree  $n$  and associated with the parameters  $\alpha - \frac{1}{2}, \alpha - \frac{1}{2}$ ;  $L_n(x)$  is the  $n^{\text{th}}$ -degree Legendre polynomial,  $T_n(x)$  and  $U_n(x)$  are the  $n^{\text{th}}$ -degrees Chebyshev polynomials of the

## Chapter 6

---

first and second kinds, respectively. The Gegenbauer polynomials constrained by standardization (6.19) are generated using the three-term recurrence relation

$$(j + 2\alpha)C_{j+1}^{(\alpha)}(x) = 2(j + \alpha)x C_j^{(\alpha)}(x) - j C_{j-1}^{(\alpha)}(x), \quad j \geq 1, \quad (6.20)$$

starting from  $C_0^{(\alpha)}(x) = 1$ ;  $C_1^{(\alpha)}(x) = x$ . At the special values  $\pm 1$ , the Gegenbauer polynomials satisfy the relation

$$C_n^{(\alpha)}(\pm 1) = (\pm 1)^n. \quad (6.21)$$

The Gegenbauer polynomials satisfy the orthogonality relation (Elgindy and Smith-Miles, 2013b)

$$\int_{-1}^1 (1 - x^2)^{\alpha - \frac{1}{2}} C_m^{(\alpha)}(x) C_n^{(\alpha)}(x) dx = \lambda_n^{(\alpha)} \delta_{mn}, \quad (6.22)$$

where

$$\lambda_n^{(\alpha)} = \frac{2^{2\alpha-1} n! \Gamma^2(\alpha + \frac{1}{2})}{(n + \alpha) \Gamma(n + 2\alpha)}; \quad (6.23)$$

$\delta_{mn}$  is the Kronecker delta function. The interested reader may further pursue more information about the class of the Gegenbauer polynomials in many useful textbooks and monographs, cf. (Abramowitz and Stegun, 1965; Szegő, 1975), for instances.

### 6.5.2 The Optimal Gegenbauer Quadrature and Definite Integrals Approximations

The method of establishing an optimal Gegenbauer quadrature was recently outlined by Elgindy and Smith-Miles (2013b) in the following theorem:

**Theorem 6.5.1** (The optimal Gegenbauer quadrature). *Let*

$$S_{N,M} = \{z_{i,k} | C_{M+1}^{(\alpha_i^*)}(z_{i,k}) = 0, i = 0, \dots, N; k = 0, \dots, M\}, \quad (6.24)$$

*be the generalized/adjoint GG points set, where  $\alpha_i^*$  are the optimal Gegenbauer parameters in the sense that*

$$\alpha_i^* = \underset{\alpha > -1/2}{\operatorname{argmin}} \eta_{i,M}^2(\alpha), \quad (6.25)$$

$$\eta_{i,M}(\alpha) = \int_{-1}^{x_i} C_{M+1}^{(\alpha)}(x) dx / K_{M+1}^{(\alpha)}; \quad (6.26)$$

$$K_{M+1}^{(\alpha)} = 2^M \frac{\Gamma(M + \alpha + 1) \Gamma(2\alpha + 1)}{\Gamma(M + 2\alpha + 1) \Gamma(\alpha + 1)}. \quad (6.27)$$

## Chapter 6

---

Moreover, let  $f(x) \in C^\infty[-1, 1]$  be approximated by the Gegenbauer polynomials expansion series such that the Gegenbauer coefficients are computed by interpolating the function  $f(x)$  at the adjoint GG points  $z_{i,k} \in S_{N,M}$ . Then there exist a matrix  $P^{(1)} = (p_{ij}^{(1)})$ ,  $i = 0, \dots, N$ ;  $j = 0, \dots, M$ ; some numbers  $\xi_i \in [-1, 1]$  satisfying

$$\int_{-1}^{x_i} f(x) dx = \sum_{k=0}^M p_{ik}^{(1)}(\alpha_i^*) f(z_{i,k}) + E_M^{(\alpha_i^*)}(x_i, \xi_i), \quad (6.28)$$

where

$$p_{ik}^{(1)}(\alpha_i^*) = \sum_{j=0}^M (\lambda_j^{(\alpha_i^*)})^{-1} \omega_k^{(\alpha_i^*)} C_j^{(\alpha_i^*)}(z_{i,k}) \int_{-1}^{x_i} C_j^{(\alpha_i^*)}(x) dx, \quad (6.29)$$

$$(\omega_k^{(\alpha_i^*)})^{-1} = \sum_{j=0}^M (\lambda_j^{(\alpha_i^*)})^{-1} (C_j^{(\alpha_i^*)}(z_{i,k}))^2, \quad (6.30)$$

$$\lambda_j^{(\alpha_i^*)} = \frac{2^{2\alpha_i^* - 1} j! \Gamma^2(\alpha_i^* + \frac{1}{2})}{(j + \alpha_i^*) \Gamma(j + 2\alpha_i^*)}; \quad (6.31)$$

$$E_M^{(\alpha_i^*)}(x_i, \xi_i) = \frac{f^{(M+1)}(\xi_i)}{(M+1)!} \eta_{i,M}(\alpha_i^*). \quad (6.32)$$

*Proof.* See (Elgindy and Smith-Miles, 2013b).  $\square$

The matrix  $P^{(1)}$  is the 1st-order optimal Gegenbauer integration matrix, and is referred to by the optimal P-matrix. To describe the approximations of the definite integrals  $\int_{-1}^{x_i} f(x) dx$  of  $f(x)$  in matrix form using the P-matrix, let  $P^{(1)} = (P_0^{(1)} P_1^{(1)} \dots P_N^{(1)})^T$ ,  $P_i^{(1)} = (p_{i,0}^{(1)}, p_{i,1}^{(1)}, \dots, p_{i,M}^{(1)})$ ;  $i = 0, \dots, N$ . Let also  $V$  be a matrix of size  $(M+1) \times (N+1)$  defined as  $V = (V_0 V_1 \dots V_N)$ ,  $V_i = (f(z_{i,0}), f(z_{i,1}), \dots, f(z_{i,M}))^T$ ,  $i = 0, \dots, N$ ;  $f(z_{ij})$  is the function  $f$  calculated at the adjoint GG nodes  $z_{i,j} \in S_{N,M}$ . Then the approximations of the definite integrals  $\int_{-1}^{x_i} f(x) dx$  of  $f(x)$  using the P-matrix are given by

$$\left( \int_{-1}^{x_0} f(x) dx, \int_{-1}^{x_1} f(x) dx, \dots, \int_{-1}^{x_N} f(x) dx \right)^T \approx P^{(1)} \circ V^T, \quad (6.33)$$

where  $\circ$  is the Hadamard product, with the elements of  $P^{(1)} \circ V^T$  given by

$$(P^{(1)} \circ V^T)_i = P_i^{(1)} \cdot V_i = \sum_{j=0}^M p_{i,j}^{(1)} f(z_{i,j}), \quad i = 0, \dots, N. \quad (6.34)$$

The reader may consult (Elgindy and Smith-Miles, 2013b,c) for further information on the developed Gegenbauer quadrature and its applications.

## PART B: Suggested Declaration for Thesis Chapter

Monash University

### Declaration for Thesis Chapter 7

#### Declaration by candidate

In the case of Chapter 7, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
The author of the key ideas, programming codes, organization, development, and writing up of the article	90

The following co-authors contributed to the work. Co-authors who are students at Monash University must also indicate the extent of their contribution in percentage terms:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Kate Smith-Miles	Provided valuable comments and aided proofreading	

Candidate's  
Signature

	Date 11/05/2013
--	--------------------

#### Declaration by co-authors

The undersigned hereby certify that:

- (1) the above declaration correctly reflects the nature and extent of the candidate's contribution to this work, and the nature of the contribution of each of the co-authors.
- (2) they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
- (3) they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (4) there are no other authors of the publication according to these criteria;
- (5) potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
- (6) the original data are stored at the following location(s) and will be held for at least five years from the date indicated below:

Location(s)

Signature 1

	Date 14/5/13
---	--------------

.....

This page is intentionally left blank

## *Chapter 7*

# *Fast, Accurate, and Small-Scale Direct Trajectory Optimization Using a Gegenbauer Transcription Method*

**Chapter 7 is based on the published article Elgindy, K. T., Smith-Miles, K. A., 15 October 2013. Fast, accurate, and small-scale direct trajectory optimization using a Gegenbauer transcription method. Journal of Computational and Applied Mathematics 251 (0), 93–116.**

**Abstract.** *This chapter reports a novel direct Gegenbauer (ultraspherical) transcription method (GTM) for solving continuous-time optimal control (OC) problems (CTOCPs) with linear/nonlinear dynamics and path constraints. In (Elgindy et al., 2012), we presented a GTM for solving nonlinear CTOCPs directly for the state and the control variables, and the method was tailored to find the best path for an unmanned aerial vehicle mobilizing in a stationary risk environment. This chapter extends the GTM to deal further with problems including higher-order time derivatives of the states by solving the CTOCP directly for the control  $u(t)$  and the highest-order time derivative  $x^{(N)}(t)$ ,  $N \in \mathbb{Z}^+$ . The state vector and its derivatives up to the  $(N - 1)^{\text{th}}$ -order derivative can then be stably recovered by successive integration. Moreover, we present our solution method for solving linear–quadratic regulator (LQR) problems as we aim to cover a wider collection of CTOCPs with the concrete aim of comparing the efficiency of the current work with other classical discretization methods in the literature. The proposed numerical scheme fully parameterizes the state and the control variables using Gegenbauer expansion series. For problems with various order time derivatives of the state variables arising in the cost function, dynamical system, or path/terminal constraints, the GTM seeks to fully parameterize the control variables and the highest-order time derivatives of the state variables. The time horizon is mapped onto the closed interval  $[0, 1]$ . The dynamical system characterized by differential equations is transformed into its integral formulation through direct integration. The resulting problem on the finite interval is then transcribed into a nonlinear programming (NLP) problem through collocation at the Gegenbauer-Gauss (GG) points. The integral operations are approximated by optimal Gegenbauer quadratures in a certain optimality sense. The reduced NLP problem is solved in the Gegenbauer spectral space, and the state and the control variables are approximated on the entire finite horizon. The proposed method achieves discrete solutions exhibiting exponential convergence using relatively small-scale number of collocation points. The advantages of the proposed direct GTM over other traditional discretization methods are shown through four well-studied OC test examples. The present work is a major breakthrough in the area of computational OC theory as it delivers significantly more accurate solutions using considerably smaller numbers of collocation points, states and controls expansion terms. Moreover, the GTM produces very small-scale NLP problems, which can be solved very*

*quickly using the modern NLP software.*

**Keyword.** *Direct optimization methods; Gegenbauer collocation; Gegenbauer integration matrix; Gegenbauer polynomials; Optimal control; Spectral methods.*

References are considered at the end of the thesis.

# Chapter 7

## Fast, Accurate, and Small-Scale Direct Trajectory Optimization Using a Gegenbauer Transcription Method

### 7.1 Introduction

The principle goal of optimal control (OC) theory is to determine the control which causes a system to meet a set of physical constraints while optimizing some performance criterion. A closed form expression of the OC is usually out of reach, and classical solution methods such as the calculus of variations, dynamic programming, and Pontryagin's maximum/minimum principle can only provide the analytical OC in very special cases. Fortunately, the immense evolution today in the fields of numerical analysis and approximation theory, and the increasing developments in digital computers, have allowed the treatment of complex OC problems by sophisticated numerical methods (Elgindy et al., 2012). Among the available numerical schemes, direct optimization methods, which transcribe the infinite-dimensional continuous-time OC problem (CTOCP) into a finite-dimensional parameter nonlinear programming (NLP) problem, have become the ideal methods of choice nowadays (Gong et al., 2006a, 2008; Hesthaven et al., 2007), and are well-suited for solving intricate OC problems; cf. (Benson et al., 2006; Betts, 2009; Chen et al., 2011; Elnagar et al., 1995; Elnagar and Razzaghi, 1997; Fahroo and Ross, 2002, 2008; Garg et al., 2011a,b; Gong et al., 2006a; Hull, 1997; Jaddu, 2002; Kang et al., 2007, 2008; Razzaghi and Elnagar, 1993; Stryk, 1993; Vlassenbroeck and Dooren, 1988; Williams, 2004) and the references therein.

## Chapter 7

---

A critical stage in the transcription of a CTOCP by a direct optimization method manifests in the discretization of the dynamics. This can be carried out using finite difference schemes, such as Euler's method and Runge-Kutta methods, finite element methods, piecewise-continuous polynomials such as linear splines, cubic splines, and B-spline methods, wavelets methods such as Walsh-wavelets and Haar wavelets methods, block pulse function methods, etc.; cf. (Becker and Vexler, 2007; Chen and Lu, 2010; Dontchev and Hager, 1997; Dontchev et al., 2000; Glabisz, 2004; Hargraves and Paris, 1987; Hsiao, 1997; Hwang et al., 1986; Kadalbajoo and Yadaw, 2008; Kaya and Martínez, 2007; Kiparissides and Georgiou, 1987; Lang and Xu, 2012; Liu et al., 2004; Liu and Yan, 2001; Pytlak, 1998; Schwartz and Polak, 1996; Stryk, 1993; Xing et al., 2010). However, a common feature in all of these numerical methods is that they usually experience an explosion in the number of variables if high orders of accuracy are sought except for very special cases, where the control is of a bang-bang control type (Kaya, 2010). This is due to the finite-order convergence rates associated with these methods (Weideman and Reddy, 2000). Spectral methods, among the available discretization methods in the literature, are memory minimizing, provide Eulerian-like simplicity, produce global solutions and rapid convergence, and are so accurate for problems exhibiting smooth solutions to the extent that they are often used in cases when "nearly exact numerical solutions are sought" (Barranco and Marcus, 2006; Cushman-Roisin and Beckers, 2011; Gardner et al., 1989; Gong et al., 2007; Gottlieb and Orszag, 1977; Zang et al., 1982). All of these advantages place the spectral methods at the front of the available numerical methods for solving ordinary and partial differential equations, eigenvalue problems, OC problems, and in many other applications exhibiting sufficiently differentiable solutions; cf. (Boyd, 2001; Elgindy, 2009; Elgindy and Hedar, 2008; Elnagar et al., 1995; Fahroo and Ross, 2002; Gong et al., 2006a; Mason and Handscomb, 2003; Quarteroni and Valli, 1994; Ross and Fahroo, 2003). Collocation/pseudospectral methods is a distinguished class of spectral methods, which have emerged as important and popular computational methods for the numerical solution of OC problems in the last two decades; cf. (Benson, 2004; Benson et al., 2006; Elnagar et al., 1995; Elnagar, 1997; Elnagar and Razzaghi, 1997; Fahroo and Ross, 2008; Garg et al., 2011b, 2010; Huntington, 2007; Rao et al., 2010; Williams, 2004). They are already well established in the works of Canuto et al. (1988, 2006, 2007); Gottlieb and Orszag (1977); Trefethen (2000), and in many other works in the literature. Their universal application in many areas is largely due to their greater simplicity and computational efficiency compared to other spectral methods, namely, Galerkin and tau methods (Gottlieb and Orszag, 1977). Perhaps one of the significant applications of spectral collocation methods that has received wide publicity recently was in generating real time trajectories for a NASA spacecraft maneuver (Kang and Bedrossian, 2007).

## Chapter 7

---

Another central element in the numerical discretization of a CTOCP lies in the accurate translation of the integral operations into precise algebraic expressions to produce an accurate discrete analog of the original CTOCP. The integral operations may occur in the cost function, the dynamical system, or in the path/terminal constraints. Even for dynamical systems characterized by a set of differential equations, it has been shown by El-Gendi (1969); Elbarbary (2007); Elgindy (2009); Elgindy and Smith-Miles (2013c); Elgindy et al. (2012); Gonzales et al. (1997); Greengard (1991); Greengard and Rokhlin (1991) that a more practical and robust numerical scheme can be established by recasting the dynamical system into its integral formulation. In the latter case, the integration points  $\{x_i\}_{i=0}^N$  of the definite integrals  $\int_{-1}^{x_i} f_j(x)dx, j \in \mathbb{Z}^+$ , of some integrand functions  $f_j(x)$ , arise naturally as the very same collocation points employed in the spectral integration method; cf. (Elgindy and Smith-Miles, 2013b,c; Elgindy et al., 2012), for instance. While traditional spectral methods demand that the number of spectral expansion terms  $(N + 1)$  required for the construction of the spectral operational matrix of differentiation/integration be exactly the same as the number of collocation points; cf. (El-Gendi, 1969; Elbarbary, 2007; Elnagar, 1997; Fornberg, 1990; Ghoreishi and Hosseini, 2004; Gong et al., 2009; Paraskevopoulos, 1983; Ross and Fahroo, 2002; Weideman and Reddy, 2000), we broke the parity restriction on the number of expansion terms and the number of integration/collocation points, and established an optimal rectangular Gegenbauer (ultraspherical) integration matrix, where the choice of the number of Gegenbauer expansion terms  $(M + 1)$  is completely free; cf. (Elgindy and Smith-Miles, 2013b, Theorem 2.2 in pg. 86). This novel approach in the constitution of the operational matrix of integration and its associated numerical quadrature gives preference to the direct Gegenbauer collocation method endowed with the optimal Gegenbauer quadrature over traditional spectral collocation methods from two perspectives (Elgindy and Smith-Miles, 2013b): (i) For any small number of collocation points  $(N + 1)$ , the Gegenbauer collocation method can boost the precision of the approximate solutions by increasing the number of optimal Gegenbauer quadrature expansion terms  $(M + 1)$  without increasing the value of  $N$ . Consequently, one can achieve higher-order approximations to the solutions of complex CTOCPs without increasing the number of collocation points. The reader may consult our recent chapter (Elgindy and Smith-Miles, 2013c) for clear examples highlighting the significance of this result. (ii) For any large number of collocation points  $(N + 1)$ , the Gegenbauer collocation method can produce very precise approximations to the smooth solutions of the CTOCP in a short time by restricting the value of  $M$  to accept only small values, and deterring it from growing up linearly with the number  $N$ .

In (Elgindy et al., 2012), we presented a Gegenbauer transcription method (GTM) for solving nonlinear CTOCPs directly for the states and the controls.

## Chapter 7

---

We focused our attention on the application of the optimal Gegenbauer quadrature to find the best path for an unmanned aerial vehicle mobilizing in a stationary risk environment. Our comparisons with the variational technique of Miller et al. (2011) showed the advantages of the GTM over classical variational methods in many aspects. Our goal in this chapter is to extend the method presented in (Elgindy et al., 2012) to handle CTOCPs including higher-order time derivatives of the states by solving the CTOCP directly for the control  $u(t)$  and the highest-order time derivative of the state,  $x^{(N)}(t)$ ,  $N \in \mathbb{Z}^+$ . To this end, we introduce a substitution for the highest-order time derivative  $x^{(N)}(t)$ , and recover the state vector and its derivatives up to the  $(N-1)^{\text{th}}$ -order derivative stably by successive integration. This key idea provides the luxury of working in a full integration environment, enjoying the well-stability of the integral operators. We shall consider also the solution of the linear-quadratic regulator (LQR) problem characterized by a linear time-invariant dynamical system as we intend to cover a wider collection of problems with the concrete aim of comparing the performance of the GTM with its rivals in the class of direct orthogonal collocation/pseudospectral methods. We highlight the degree of robustness, simplicity, accuracy, economy in calculations, and speed of the GTM compared to other conventional methods in the area of computational OC theory. Furthermore, we endeavor to establish a high-order numerical scheme which results in a NLP problem with considerably lower-dimensional space to facilitate the task of the NLP solver and reduce the calculation time.

The proposed method converts the CTOCP into a NLP problem through a Gegenbauer collocation scheme based on GG points. The solution technique converts the dynamical system of the differential equations form into integral equations through direct integration. The state and the control variables are fully parameterized and approximated by truncated Gegenbauer expansion series with unknown Gegenbauer collocation coefficients. The integral operations are approximated by the optimal Gegenbauer quadratures developed in (Elgindy and Smith-Miles, 2013b). The proposed technique reduces the cost function, the dynamics, and the constraints into systems of algebraic equations, and thus greatly simplifies the problem. In this manner, the infinite-dimensional CTOCP is transcribed into a finite-dimensional parameter NLP problem, which can be solved for the Gegenbauer collocation coefficients using the powerful and well-developed NLP software and computer codes. For problems with higher-order time derivatives of the states arising in the performance index, dynamics, or path/terminal constraints, we parameterize the control and the highest-order time derivative of the state. We restrict ourselves to developing algorithms for solving CTOCPs governed by ordinary differential equations. The CTOCPs governed by integro-differential equations can be solved similarly by recasting the dynamics into its integral formulation. Moreover, the CTOCPs governed by integral equations can

## Chapter 7

---

be discretized directly by the Gegenbauer quadratures to approximate the integral operators and transform the integral equations into algebraic equations.

The remaining of the chapter is organized as follows: In Section 7.2 we present the CTOCP statements considered in this chapter. The proposed GTM for the solution of the considered CTOCPs is introduced in Section 7.3. We highlight the convergence and some advantages of the GTM in Section 7.4. Four numerical experiments well-studied in the literature are presented in Section 7.5 to demonstrate the efficiency, robustness, and the spectral accuracy of the proposed method. Section 7.6 is devoted to a discussion on the proposed GTM showing its strengths over the traditional discretization methods followed by some concluding remarks and future works associated with the current GTM. Further brief information on the Gegenbauer polynomials is provided in the appendix.

### 7.2 The CTOCPs statements

In the following, we present the CTOCP statements considered in this chapter:

(I) Consider the following LQR problem  $\mathcal{P}_1$ :

$$\text{minimize } J(u(t)) = \frac{1}{2}(x(1)^T S x(1)) + \frac{1}{2} \int_0^1 (x(t)^T A x(t) + u(t)^T B u(t)) dt, \quad (7.1a)$$

$$\text{subject to } \dot{x}(t) = D x(t) + E u(t), \quad (7.1b)$$

$$x(0) = x^0, \quad (7.1c)$$

where  $[0, 1]$  is the time interval of interest,  $x \in \mathbb{R}^n$  is the state vector,  $\dot{x} \in \mathbb{R}^n$  is the vector of first-order time derivatives of the states,  $u \in \mathbb{R}^m$  is the control vector,  $J$  is the cost function to be minimized,  $S$  and  $A$  are constant positive semidefinite matrices,  $B$  is a constant positive definite matrix;  $D$  and  $E$  are constant matrices. Here the LQR problem is characterized by the linear time-invariant dynamical system (7.1b) and the initial state condition (7.1c). The optimal LQR problem is to determine the OC policy  $u^*(t)$  on the time horizon  $[0, 1]$  which meets Constraints (7.1b) & (7.1c) while minimizing the quadratic cost functional (7.1a). The corresponding optimal state trajectory is denoted by  $x^*(t)$ .

(II) The second CTOCP we are concerned with arises due to the presence of some high-order derivatives of the state variables in the dynamical system, the cost function, the path constraints, or the terminal constraints. For instance, consider the following nonlinear CTOCP with fixed final time,

## Chapter 7

---

some mixed path and terminal inequality constraints  $\mathcal{P}_2$ :

$$\text{minimize } J(u(t)) = \Phi(x(1)) + \int_0^1 \mathcal{L}(x(t), \dot{x}(t), \dots, x^{(m')}(t), u(t), t) dt, \quad (7.2a)$$

$$\text{subject to } x^{(m')}(t) = f(x(t), \dot{x}(t), \dots, x^{(m'-1)}(t), u(t), t), \quad (7.2b)$$

$$x(0) = \mathcal{K}_0, \dot{x}(0) = \mathcal{K}_1, \dots, x^{(m'-1)}(0) = \mathcal{K}_{m'-1}, \quad (7.2c)$$

$$\psi_i(x(t), \dot{x}(t), \dots, x^{(m'-1)}(t), x^{(m')}(t), u(t), t) \leq 0, \quad i = 0, \dots, \ell, \quad (7.2d)$$

$$\phi(x(1), \dot{x}(1), \dots, x^{(m'-1)}(1), x^{(m')}(1), u(1)) \leq 0, \quad m' \in \mathbb{Z}^+. \quad (7.2e)$$

This problem is a Bolza problem, where  $x^{(k)} \in \mathbb{R}^n$  is the vector of  $k^{\text{th}}$ -order time derivatives of the states for each  $k = 1, \dots, m'$ ;  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$  is the terminal cost function,  $\mathcal{L} : \mathbb{R}^n \times \dots \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}$  is the Lagrangian function,  $f : \mathbb{R}^n \times \dots \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^n$  is a nonlinear vector field,  $\psi_i : \mathbb{R}^n \times \dots \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}$  is a mixed inequality constraint on the states, their derivatives, and the controls for each  $i$ ,  $\phi : \mathbb{R}^n \times \dots \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$  is a mixed terminal inequality constraint;  $\mathcal{K}_0, \dots, \mathcal{K}_{m'-1}$  are the initial conditions of the nonlinear system dynamics (7.2b). Here it is assumed that  $\Phi$ ,  $\mathcal{L}$ , and each system function  $f_i$  are continuously differentiable with respect to  $x$ ;  $\mathcal{L}$  and  $f_i$  are continuous with respect to  $u$ .

For both problems  $\mathcal{P}_1$  &  $\mathcal{P}_2$ , we shall assume that the dynamical system has a unique state trajectory  $x(t)$  for any admissible control trajectory  $u(t)$ . Notice that a CTOCP defined over the physical time domain  $[t_0, t_f]$  can be reformulated into the form of Problems  $\mathcal{P}_1$  &  $\mathcal{P}_2$  using the strict change of variable  $t = (\tau - t_0)/(t_f - t_0)$ , where  $\tau \in [t_0, t_f]$ ,  $t_0$  and  $t_f$  are the initial and final times, respectively. In the next section, we shall describe the GTM for the numerical solution of Problems  $\mathcal{P}_1$  &  $\mathcal{P}_2$  based on GG collocation.

### 7.3 The GTM

To approximate the system dynamics in Problems  $\mathcal{P}_1$  &  $\mathcal{P}_2$ , it is necessary to find expressions for the derivatives of the state variables at the collocation points. This can be accomplished through spectral differentiation matrices (SDMs). Nonetheless, numerical differentiation is in principle an ill-posed problem (Liu et al., 2011), and SDMs are known to be severely ill-conditioned (Driscoll, 2010; Elbarbary, 2006, 2007; Funaro, 1987). Therefore the implementation of SDMs causes degradation in the observed precision (Driscoll, 2010; Tang and Trummer, 1996).

## Chapter 7

---

In fact, the condition number of the  $N^{\text{th}}$ -order SDM is typically of order  $O(N^{2k})$ , where  $k$  is the order of the derivative of the solution function (Hesthaven, 2000). This fact casts its shadow on imposing strict stability requirements when the SDM is associated with a numerical ODE solver to solve a time-dependent PDE, for instance (Kong and Rokhlin, 2012). Moreover, the time step restrictions can be more severe than those predicted by the standard stability theory (Trefethen, 1988; Trefethen and Trummer, 1987). To quote (Elgindy et al., 2012) ‘for higher-order SDMs, the ill-conditioning becomes very critical to the extent that developing efficient preconditioners is extremely crucial.’ Another useful approach is to transform the dynamical system into its integral formulation, where the state and the control variables are approximated by truncated spectral expansion series while the integral operations are approximated by spectral integration matrices (SIMs). This numerical technique is generally well-behaved as the SIMs are known to be well-conditioned operators (Elbarbary, 2006, 2007; Elgindy, 2009; Elgindy et al., 2012; Greengard, 1991; Lundbladh et al., 1992), and their well-conditioning is essentially unaffected for increasing number of grid points (Elgindy, 2009). For two-point boundary value problems, for instance, Greengard and Rokhlin (1991) showed that the integral equation formulation is insensitive to boundary layers, insensitive to end-point singularities, and leads to small condition numbers while achieving high computational efficiency. Furthermore, to quote (Elgindy et al., 2012) ‘the use of integration for constructing the spectral approximations improves the rate of convergence of the spectral interpolants, and allows the multiple boundary conditions to be incorporated more efficiently.’ These useful features, in addition to the promising results obtained by Elgindy and Smith-Miles (2013b,c); Elgindy et al. (2012), motivate us to apply a Gegenbauer collocation integration scheme for discretizing the dynamical systems of the underlying CTOCPs.

To efficiently implement the Gegenbauer collocation integration method, one needs an accurate and robust numerical quadrature to perfectly translate the integral operations into their algebraic expressions analog. In (Elgindy and Smith-Miles, 2013b), we showed that an optimal Gegenbauer quadrature can be constituted by combining the strengths of the Chebyshev, Legendre, and Gegenbauer polynomials in a unique numerical quadrature through a unified approach. In particular, the developed optimal quadrature employs the Gegenbauer polynomials to achieve rapid convergence rates of the quadrature in the small/medium range of the spectral expansion terms. For a large-scale number of expansion terms, the numerical quadrature possesses the luxury of converging to the optimal Chebyshev and Legendre quadratures in the  $L^\infty$ -norm and  $L^2$ -norm, respectively. The key idea in our work is to approximate the definite integrals  $\int_{-1}^{x_i} f(x)dx, 0 \leq i \leq N$ , of a given smooth function  $f(x)$  by constructing the Gegenbauer quadrature through discretizations/interpolations at some optimal

## Chapter 7

---

sets of the Gegenbauer-Gauss (GG) points  $\{z_{i,j}\}_{j=0}^M$ ,  $M \in \mathbb{Z}^+$ , which are determined by satisfying a certain optimality measure to minimize the quadrature error and produce faster convergence rates. These optimal GG interpolation points are called the adjoint GG points, and they usually differ than the given integration points  $\{x_i\}_{i=0}^N$ . In the following two sections, we shall describe how to implement the optimal Gegenbauer quadrature within the framework of a Gegenbauer collocation integration scheme for the solution of Problems  $\mathcal{P}_1$  &  $\mathcal{P}_2$ .

### 7.3.1 Solving Problem $\mathcal{P}_1$ using the GTM

Integrating Equation (7.1b) and using the initial condition (7.1c) recast the dynamical system into its integral formulation given by

$$x(t) = D \int_0^t x(\tau) d\tau + E \int_0^t u(\tau) d\tau + x^0. \quad (7.3)$$

To transcribe the CTOCP  $\mathcal{P}_1$ , we expand the state and the control variables by the Gegenbauer expansion series

$$x_r(t) \approx \sum_{k=0}^L a_{rk} C_k^{(\alpha)}(t), \quad r = 1, \dots, n, \quad (7.4)$$

$$u_s(t) \approx \sum_{k=0}^M b_{sk} C_k^{(\alpha)}(t), \quad s = 1, \dots, m, \quad (7.5)$$

and collocate at the GG nodes  $t_i \in S_N^{(\alpha)}$  defined by Equation (7.A.4), since they have the desirable distribution property of clustering around the endpoints of the interval; thus avoiding the Runge phenomenon (Hesthaven et al., 2007; Trefethen, 2000). Following the mathematical convention introduced in (Elgindy and Smith-Miles, 2013b), let

$$S_{N,M_P} = \{z_{i,k} | C_{M_P+1}^{(\alpha_i^*)}(z_{i,k}) = 0, i = 0, \dots, N; k = 0, \dots, M_P\}, \quad M_P \in \mathbb{Z}^+, \quad (7.6)$$

be the adjoint set of the GG points, where

$$\alpha_i^* = \operatorname{argmin}_{\alpha > -1/2} \eta_{i,M_P}^2(\alpha), \quad (7.7)$$

$$\eta_{i,M_P}(\alpha) = \int_{-1}^{x_i} C_{M_P+1}^{(\alpha)}(x) dx / K_{M_P+1}^{(\alpha)}; \quad (7.8)$$

$$K_{M_P+1}^{(\alpha)} = 2^{M_P} \frac{\Gamma(M_P + \alpha + 1) \Gamma(2\alpha + 1)}{\Gamma(M_P + 2\alpha + 1) \Gamma(\alpha + 1)}. \quad (7.9)$$

## Chapter 7

Let also  $P^{(1)} = (P_0^{(1)} P_1^{(1)} \dots P_N^{(1)})^T$  be the 1st-order optimal Gegenbauer operational matrix of integration referred to by the optimal P-matrix, where  $P_i^{(1)} = (p_{i,0}^{(1)}, p_{i,1}^{(1)}, \dots, p_{i,M_P}^{(1)})$ ,  $0 \leq i \leq N$ , and  $p_{ik}^{(1)}$ ,  $0 \leq i \leq N$ ,  $0 \leq k \leq M_P$ , are the elements of the optimal P-matrix as defined by Equation (2.20) in (Elgindy and Smith-Miles, 2013b). To simplify the calculations, we shall introduce the following mathematical notations:  $t_{N+1} = 1$ ,  $\hat{e}_l \in \mathbb{R}^{l+1} : (\hat{e}_l)_k = 1$ ,  $z_{i'j} \in S_{N+1, M_P}$ ,  $a = (a_1, \dots, a_n)^T$ ,  $b = (b_1, \dots, b_m)^T$ ,  $a_r = (a_{r0}, \dots, a_{rL})^T$ ,  $b_s = (b_{s0}, \dots, b_{sM})^T$ ,  $\xi_l^{(\alpha)} \in \mathbb{R}^{(N+1) \times (l+1)} : (\xi_l^{(\alpha)})_{ik} = (C_k^{(\alpha)}(t_i))$ ,  $\hat{\xi}_{li}^{(\alpha)T} \in \mathbb{R}^{l+1} : (\hat{\xi}_{li}^{(\alpha)})_k = (\xi_l^{(\alpha)})_{ik}$ ,  $\zeta_l^{(\alpha)} \in \mathbb{R}^{(N+2) \times (M_P+1) \times (l+1)} : (\zeta_l^{(\alpha)})_{i'jk} = (C_k^{(\alpha)}(z_{i'j}))$ ,  $\hat{\zeta}_{li'j}^{(\alpha)T} \in \mathbb{R}^{l+1} : (\hat{\zeta}_{li'j}^{(\alpha)})_k = (\zeta_l^{(\alpha)})_{i'jk}$ ,  $\bar{\zeta}_{li'k}^{(\alpha)} \in \mathbb{R}^{M_P+1} : (\bar{\zeta}_{li'k}^{(\alpha)})_j = (\zeta_l^{(\alpha)})_{i'jk}$ ,  $\chi_{li'}^{(\alpha)} \in \mathbb{R}^{(M_P+1) \times (l+1)} : (\chi_{li'}^{(\alpha)})_{jk} = (\zeta_l^{(\alpha)})_{i'jk}$ ,  $\wp_{lq}^{(\alpha)} \in \mathbb{R}^{(N+1) \times (l+1)} : (\wp_{lq}^{(\alpha)})_{ik} = P_i^{(q)} \bar{\zeta}_{lik}^{(\alpha)}$ ,  $\hat{\wp}_{liqi}^{(\alpha)T} \in \mathbb{R}^{l+1} : (\hat{\wp}_{liqi}^{(\alpha)})_k = (\wp_{lq}^{(\alpha)})_{ik}$ ,  $(\hat{\wp}_{liq, N+1}^{(\alpha)})_k = P_{N+1}^{(q)} \bar{\zeta}_{li, N+1, k}^{(\alpha)}$ ,  $r = 1, \dots, n$ ;  $s = 1, \dots, m$ ;  $i = 0, \dots, N$ ;  $i' = 0, \dots, N+1$ ;  $j = 0, \dots, M_P$ ;  $k = 0, \dots, l$ ;  $l \in \mathbb{Z}^+$ . The  $q$ -fold integral of the Gegenbauer polynomials  $I_{q,k}^{(\alpha)}(t_i)$ , for the integration points set  $S_N^{(\alpha)}$ , can be approximated by the P-matrix as follows:

$$I_{q,k}^{(\alpha)}(t_i) = \int_0^{t_i} \dots \int_0^{\tau_2} C_k^{(\alpha)}(\tau_1) d\tau_1 \dots d\tau_q \approx P_i^{(q)} \bar{\zeta}_{lik}^{(\alpha)}. \quad (7.10)$$

The relation between the matrix  $(I_{q,k}^{(\alpha)}(t_i)) \in \mathbb{R}^{(N+1) \times (l+1)}$  and the  $q^{\text{th}}$ -order P-matrix is given by

$$(I_{q,k}^{(\alpha)}(t_i)) \approx \wp_{lq}^{(\alpha)}. \quad (7.11)$$

Using these notations, the state and the control vectors at the GG collocation points can be written as

$$x(t_i) \approx (I_n \otimes \hat{\xi}_{Li}^{(\alpha)})a, \quad (7.12)$$

$$u(t_i) \approx (I_m \otimes \hat{\xi}_{Mi}^{(\alpha)})b, \quad (7.13)$$

where  $I_l$  is the identity matrix of order  $l$ ;  $\otimes$  is the Kronecker product of matrices. Using Equation (7.A.5), we can show that  $x(1) = (I_n \otimes \hat{e}_L^T)a$ . Hence the discrete cost function can be represented by

$$J \approx \tilde{J}(a, b) = \frac{1}{2} \left( a^T (S \otimes \mathcal{J}_{L+1}) a + P_{N+1}^{(1)} \hat{\mathcal{L}} \right), \quad (7.14)$$

where

$$(\hat{\mathcal{L}})_j = a^T \left( A \otimes \hat{\zeta}_{L, N+1, j}^{(\alpha)T} \hat{\zeta}_{L, N+1, j}^{(\alpha)} \right) a + b^T \left( B \otimes \hat{\zeta}_{M, N+1, j}^{(\alpha)T} \hat{\zeta}_{M, N+1, j}^{(\alpha)} \right) b; \quad (7.15)$$

$\mathcal{J}_{L+1}$  is the ones matrix of order  $(L+1)$ . The discrete dynamical system becomes

$$\begin{aligned} \mathcal{H}_i(a, b) &= \left( (I_n \otimes \hat{\xi}_{Li}^{(\alpha)}) - D(I_n \otimes P_i^{(1)} \chi_{Li}^{(\alpha)}) \right) a - E(I_m \otimes P_i^{(1)} \chi_{Mi}^{(\alpha)}) b - x^0 \approx 0, \\ i &= 0, \dots, N, \end{aligned} \quad (7.16)$$

## Chapter 7

---

which can be simplified further to

$$\mathcal{H}_i(a, b) = \left( (I_n \otimes \hat{\xi}_{Li}^{(\alpha)}) - D(I_n \otimes \hat{\phi}_{Li}^{(\alpha)}) \right) a - E(I_m \otimes \hat{\phi}_{M1i}^{(\alpha)}) b - x^0 \approx 0, \quad i = 0, \dots, N. \quad (7.17)$$

Hence the GTM transforms the CTOCP  $\mathcal{P}_1$  into a constrained NLP problem of the form:

$$\text{minimize } \tilde{J}(a, b) \quad (7.18a)$$

$$\text{subject to } \mathcal{H}_i(a, b) \approx 0, \quad i = 0, \dots, N, \quad (7.18b)$$

which can be solved using standard optimization software. Notice here that the dynamical system is enforced by the GTM as equality constraints at the internal GG collocation points. Moreover, the GTM solves the CTOCP  $\mathcal{P}_1$  in the spectral space; therefore, once the approximate Gegenbauer coefficients are found, the approximation can immediately be evaluated at any time history of both the control and the state variables without invoking any interpolation method. This feature establishes the power of the proposed GTM for solving CTOCPs as the optimal state and control profiles are readily determined; moreover, it represents a clear advantage over the “classical” discretization methods such as the finite difference schemes which require a further step of interpolation to evaluate an approximation at an intermediate point.

### 7.3.2 Solving Problem $\mathcal{P}_2$ using the GTM

The nonlinearity of Problem  $\mathcal{P}_2$ , and the existence of higher-order derivatives of the state vector in the cost function, the dynamics, and the path and terminal constraints add more complexity over Problem  $\mathcal{P}_1$  both analytically and computationally. To overcome this difficulty, we introduce the following substitution:

$$x^{(m')}(t) = \mu(t), \quad (7.19)$$

for some unknown continuous vector function  $\mu(t) \in \mathbb{R}^n$ . We also define

$$\nu_q(t) = \sum_{k=1}^q \frac{\mathcal{K}_{m'-q+k-1}}{(k-1)!} t^{k-1}, \quad (7.20)$$

and denote  $\nu_q(t_i); \nu_q(z_{ij})$  by  $\nu_{qi}; \nu_{qij}$ , respectively. Then the state vector and its derivatives up to the  $(m' - 1)^{\text{th}}$ -derivative can be obtained by successive integra-

## Chapter 7

---

tions as follows:

$$x^{(m'-1)}(t) = \int_0^t \mu(\tau) d\tau + \nu_1(t) \hat{e}_{n-1}, \quad (7.21a)$$

$$x^{(m'-2)}(t) = \int_0^t \int_0^{\tau_2} \mu(\tau_1) d\tau_1 d\tau_2 + \nu_2(t) \hat{e}_{n-1}, \quad (7.21b)$$

⋮

$$x(t) = \int_0^t \dots \int_0^{\tau_2} \mu(\tau_1) d\tau_1 \dots d\tau_{m'} + \nu_{m'}(t) \hat{e}_{n-1}. \quad (7.21c)$$

Expand the unknown variables  $\mu_r(t)$  by Gegenbauer expansion series as follows:

$$\mu_r(t) \approx \sum_{k=0}^L a_{rk} C_k^{(\alpha)}(t), \quad r = 1, \dots, n, \quad (7.22)$$

and expand the control variables by the Gegenbauer expansion series (7.5). The control vector at the solution nodes  $\{t_i\}_{i=0}^N$  is approximated by (7.13) while the state vector and its derivatives are approximated by the optimal P-matrices as follows:

$$x(t_i) \approx (I_n \otimes \hat{\phi}_{Lm'i}^{(\alpha)})a + \nu_{m'i} \hat{e}_{n-1}, \quad (7.23a)$$

$$\dot{x}(t_i) \approx (I_n \otimes \hat{\phi}_{L,m'-1,i}^{(\alpha)})a + \nu_{m'-1,i} \hat{e}_{n-1}, \quad (7.23b)$$

$$\vdots \quad (7.23c)$$

$$x^{(m'-1)}(t_i) \approx (I_n \otimes \hat{\phi}_{L,1,i}^{(\alpha)})a + \nu_{1,i} \hat{e}_{n-1}; \quad (7.23d)$$

$$x^{(m')}(t_i) \approx (I_n \otimes \hat{\xi}_{Li}^{(\alpha)})a. \quad (7.23e)$$

Hence the discrete dynamical system at the solution nodes  $\{t_i\}_{i=0}^N$  is given by

$$\begin{aligned} \mathcal{H}_i(a, b) = & (I_n \otimes \hat{\xi}_{Li}^{(\alpha)})a - f \left( (I_n \otimes \hat{\phi}_{Lm'i}^{(\alpha)})a + \nu_{m'i} \hat{e}_{n-1}, (I_n \otimes \hat{\phi}_{L,m'-1,i}^{(\alpha)})a + \nu_{m'-1,i} \hat{e}_{n-1}, \right. \\ & \left. \dots, (I_n \otimes \hat{\phi}_{L,1,i}^{(\alpha)})a + \nu_{1,i} \hat{e}_{n-1}, (I_m \otimes \hat{\xi}_{Mi}^{(\alpha)})b, t_i \right) \approx 0, \quad i = 0, \dots, N. \end{aligned} \quad (7.24)$$

The discrete path and terminal inequality constraints are given by

$$\begin{aligned} c_{ij}(a, b) = & \psi_j \left( (I_n \otimes \hat{\phi}_{Lm'i}^{(\alpha)})a + \nu_{m'i} \hat{e}_{n-1}, (I_n \otimes \hat{\phi}_{L,m'-1,i}^{(\alpha)})a + \nu_{m'-1,i} \hat{e}_{n-1}, \dots, \right. \\ & \left. (I_n \otimes \hat{\phi}_{L,1,i}^{(\alpha)})a + \nu_{1,i} \hat{e}_{n-1}, (I_n \otimes \hat{\xi}_{Li}^{(\alpha)})a, (I_m \otimes \hat{\xi}_{Mi}^{(\alpha)})b, t_i \right) \leq 0, \quad i = 0, \dots, N; \\ & j = 0, \dots, \ell, \end{aligned} \quad (7.25)$$

$$\begin{aligned} c_t(a, b) = & \phi \left( (I_n \otimes \hat{\phi}_{L,m',N+1}^{(\alpha)})a + \nu_{m',N+1} \hat{e}_{n-1}, (I_n \otimes \hat{\phi}_{L,m'-1,N+1}^{(\alpha)})a + \nu_{m'-1,N+1} \hat{e}_{n-1}, \right. \\ & \left. \dots, (I_n \otimes \hat{\phi}_{L,1,N+1}^{(\alpha)})a + \nu_{1,N+1} \hat{e}_{n-1}, (I_n \otimes \hat{e}_L^T)a, (I_m \otimes \hat{e}_M^T)b \right) \leq 0, \end{aligned} \quad (7.26)$$

## Chapter 7

---

respectively. To approximate the cost function, let

$$S_{M_P, M_{\bar{P}}} = \{\bar{z}_{jl} : C_{M_{\bar{P}+1}}^{(\bar{\alpha}_j^*)}(\bar{z}_{jl}) = 0, j = 0, \dots, M_P; l = 0, \dots, M_{\bar{P}}\}, \quad (7.27)$$

be the adjoint GG points set, for some  $M_{\bar{P}} \in \mathbb{Z}^+$ . Furthermore, let  $\bar{P}^{(q)} = (\bar{p}_{jl}^{(q)})$  be the  $q^{\text{th}}$ -order P-matrix required for the evaluation of the  $q$ -fold integral of the Gegenbauer polynomials, for the integration points set  $\mathcal{S} = \{z_{N+1,j}\}_{j=0}^{M_P}$ . Define  $\bar{P}_j^{(q)} = (\bar{p}_{j0}^{(q)}, \dots, \bar{p}_{jM_{\bar{P}}}^{(q)})$ ,  $\Lambda_l^{(\alpha)} \in \mathbb{R}^{(M_P+1) \times (M_{\bar{P}+1}) \times (l+1)} : (\Lambda_l^{(\alpha)})_{j sk} = (C_k^{(\alpha)}(\bar{z}_{js}))$ ;  $\bar{\Lambda}_{ljk}^{(\alpha)} \in \mathbb{R}^{M_{\bar{P}+1}} : (\bar{\Lambda}_{ljk}^{(\alpha)})_s = (\Lambda_l^{(\alpha)})_{j sk}$ , then

$$I_{q,k}^{(\alpha)}(z_{N+1,j}) = \int_0^{z_{N+1,j}} \dots \int_0^{\tau_2} C_k^{(\alpha)}(\tau_1) d\tau_1 \dots \tau_q \approx \bar{P}_j^{(q)} \bar{\Lambda}_{ljk}^{(\alpha)}. \quad (7.28)$$

Define  $\theta_{lq}^{(\alpha)} \in \mathbb{R}^{(M_P+1) \times (l+1)} : (\theta_{lq}^{(\alpha)})_{jk} = \bar{P}_j^{(q)} \bar{\Lambda}_{ljk}^{(\alpha)}$ ;  $\bar{\theta}_{lqj}^{(\alpha)T} \in \mathbb{R}^{l+1} : (\bar{\theta}_{lqj}^{(\alpha)})_k = (\theta_{lq}^{(\alpha)})_{jk}$ , then

$$(I_{q,k}^{(\alpha)}(z_{N+1,j})) \approx \theta_{lq}^{(\alpha)}, \quad (7.29)$$

where  $(I_{q,k}^{(\alpha)}(z_{N+1,j})) \in \mathbb{R}^{(M_P+1) \times (l+1)}$  is the matrix of the  $q$ -fold integral of the Gegenbauer polynomials for the integration points set  $\mathcal{S}$ . Equation (7.29) gives the relation between  $(I_{q,k}^{(\alpha)}(z_{N+1,j}))$  and the  $q^{\text{th}}$ -order P-matrix,  $\bar{P}^{(q)}$ . Hence the discrete cost function can be approximated by:

$$J \approx \tilde{J}(a, b) = \Phi \left( (I_n \otimes \hat{\varphi}_{L, m', N+1}^{(\alpha)})a + \nu_{m', N+1} \hat{e}_{n-1} \right) + P_{N+1}^{(1)} \hat{\mathcal{L}}, \quad (7.30a)$$

where

$$\begin{aligned} (\hat{\mathcal{L}})_j = \mathcal{L} \left( (I_n \otimes \bar{\theta}_{L, m', j}^{(\alpha)})a + \nu_{m', N+1, j} \hat{e}_{n-1}, (I_n \otimes \bar{\theta}_{L, m'-1, j}^{(\alpha)})a + \nu_{m'-1, N+1, j} \hat{e}_{n-1}, \dots, \right. \\ \left. (I_n \otimes \bar{\theta}_{L, 1, j}^{(\alpha)})a + \nu_{1, N+1, j} \hat{e}_{n-1}, (I_n \otimes \hat{\zeta}_{L, N+1, j}^{(\alpha)})a, (I_m \otimes \hat{\zeta}_{M, N+1, j}^{(\alpha)})b, z_{N+1, j} \right). \end{aligned} \quad (7.30b)$$

Eventually, the CTOCP  $\mathcal{P}_2$  is transformed into a parameter NLP problem of the following form:

$$\text{minimize } \tilde{J}(a, b), \quad (7.31a)$$

$$\text{subject to } \mathcal{H}_i(a, b) \approx 0, \quad (7.31b)$$

$$c_{ij}(a, b) \leq 0, \quad i = 0, \dots, N; j = 0, \dots, \ell; \quad (7.31c)$$

$$c_t(a, b) \leq 0, \quad (7.31d)$$

which can be solved in the spectral space using the powerful optimization methods and computer codes.

### 7.4 Properties of the GTM

Corollary 2.1 in (Elgindy and Smith-Miles, 2013b) shows that the optimal P-matrix employed in the GTM evaluates the successive integrations of the Gegenbauer polynomial of any arbitrary degree  $n$  exactly for any arbitrary sets of collocation points  $\{t_i\}_{i=0}^N$  if  $M_P \geq n$ . Therefore, the discrete dynamical system (7.17) is exact for all  $M_P \geq \max\{L, M\}$ , since the integral form of the dynamics (7.3) is linear in the states and the controls. Hence the error in the approximation of the dynamics arises only due to the round-off errors encountered during the calculations. Moreover, since the optimal P-matrix is a rectangular matrix, one can freely increase the number of its columns ( $M_P + 1$ ) while keeping the number of its rows ( $N + 1$ ) fixed. Notice here that each row of the optimal P-matrix corresponds to each of the GG collocation points  $t_i \in S_N^{(\alpha)}$  while the parameter  $M_P$  is the parameter governing the number of expansion terms in the optimal Gegenbauer quadrature. This distinctive feature of the P-matrix is extremely advantageous, since it allows for higher-order discretizations of linear/nonlinear CTOCPs without increasing the number of collocation points. Therefore, the GTM endowed with the optimal P-matrix can achieve precise approximations to the solutions of the CTOCP in short time without increasing the dimensions of the NLP problems (7.18) & (7.31). In contrast, the accuracy of typical direct pseudospectral methods is contingent upon the number of collocation points ( $N + 1$ ), which is the same as the number of spectral coefficients in the expansions of the states and the controls. Since typical spectral differentiation matrices employed in conventional direct pseudospectral methods are constructed using the collocation points employed in the discretization, the size of these square matrices must also be ( $N + 1$ ), and one usually cannot obtain higher-order approximations without increasing the size of each of these three key elements, namely, “the size of the spectral differentiation matrix, the number of collocation points, and the number of state and control expansion terms.” Eventually, to obtain comparable results to that of the present GTM, typical direct pseudospectral methods implementing full parameterization of the states and the controls require the solution of larger-scale NLP problems. We shall demonstrate these substantial results later in Section 7.5.

To analyze the convergence of the GTM, it is essential to observe the convergence of the optimal Gegenbauer quadrature associated with the P-matrix for a large number of Gegenbauer expansion terms. Theorem 2.4 in (Elgindy and Smith-Miles, 2013b) shows that the optimal Gegenbauer quadrature converges to the optimal Chebyshev quadrature in the  $L^\infty$ -norm approximation of definite integrals of smooth functions, for sufficiently large-scale number of Gegenbauer expansion terms. This attractive feature of the optimal Gegenbauer quadrature can be accomplished irrespective of the number of collocation points ( $N + 1$ ).

## Chapter 7

---

Moreover, the optimal Gegenbauer quadrature constructed via Algorithms 2.1 & 2.2 in (Elgindy and Smith-Miles, 2013b) is identical with the Legendre quadrature for large values of  $M_P$  if the approximations are sought in the  $L^2$ -norm. Hence for collocations of CTOCPs at the Chebyshev-Gauss points set  $S_N^{(0)}$  or the Legendre-Gauss points set  $S_N^{(0.5)}$ , the convergence properties of the GTM are exactly the same as those of the Chebyshev and Legendre direct orthogonal collocation methods, for large numbers of Gegenbauer expansion terms. Hence the GTM combines the strengths of the versatile Chebyshev, Legendre, and Gegenbauer polynomials in one OC solver to perform rapid and precise trajectory optimization in the sense that: (i) the Gegenbauer polynomial expansions are applied for the small/medium range of the number of spectral expansion terms to produce higher-order approximations; cf. (Elgindy and Smith-Miles, 2013b,c; Elgindy et al., 2012), and the numerical results reported in Section 7.5; (ii) the Chebyshev and Legendre polynomial expansions are applied for a large number of spectral expansion terms to produce well-conditioned and accurate approximations to the solutions of the underlying CTOCP; cf. (Gong et al., 2006a,b; Kameswaran and Biegler, 2008; Kang et al., 2005), and Example 7.5.4 in Section 7.5. The Chebyshev, Legendre, and Gegenbauer polynomials have been some of the most successful orthogonal basis polynomials by far in many applications (Fornberg, 1996), and their expansions are accurate independent of the specific boundary conditions of the solution function (Hesthaven et al., 2007).

To the best of our knowledge, the superconvergence rate of a discretization method based on collocations at Gauss points has been proven for unconstrained OC problems by Reddien (1979). Cuthrell and Biegler (1989) followed the approach of Reddien (1979), and showed equivalence between the variational optimality conditions and the Karush-Kuhn-Tucker conditions of the discretized OC problem based on Gauss collocation. Moreover, the convergence theorems and rates of the Legendre direct pseudospectral methods have been investigated in a number of articles at the Legendre-Gauss-Radau and the Legendre-Gauss-Lobatto type of collocation points; cf. (Gong et al., 2008; Kameswaran and Biegler, 2008; Kang, 2008, 2009, 2010; Kang et al., 2007; Ruths et al., 2011). However, further analysis is required to investigate the convergence of the approximate solutions obtained by general direct orthogonal collocation methods based on Gauss collocations to the solutions of the unconstrained/constrained CTOCPs.

### 7.5 Illustrative Numerical Examples

In this section we report the numerical results of the GTM for the solution of four CTOCPs well-studied in the literature. Moreover, we conducted comparisons with some other competitive OC solvers to assess the accuracy and the efficiency

## Chapter 7

---

of the present GTM. The first test example is a linear–quadratic CTOCP known as the Feldbaum problem, and represents a test model in the form of Problem  $\mathcal{P}_1$ . The second and third test examples are CTOCPs with path inequality constraints. The fourth test example is a CTOCP with a quartic Lagrangian function in the states and the control. Examples 2-4 are restated as test models in the form of Problem  $\mathcal{P}_2$ . The numerical experiments of the GTM were conducted on a personal computer having an Intel(R) Core(TM) i5 CPU with 2.53GHz speed running on a Windows 7 64-bit operating system. The reported results were obtained using the “fmincon” interior-point algorithm optimization solver provided with MATLAB V. 7.14.0.739 (R2012a). The function value termination tolerance “TolFun,” and the tolerance on the constraint violation “TolCon” were set at  $10^{-15}$ . The P-matrix was constructed using Algorithm 2.2 in (Elgindy and Smith-Miles, 2013b) with  $M_{\max} = 20$ . The reported values of  $\alpha$  were chosen among the candidate values  $-0.4 : 0.1 : 1$ .

**Example 7.5.1** (The Feldbaum problem). Find the OC  $u^*(t)$  which minimizes

$$J = \frac{1}{2} \int_0^1 (x^2(t) + u^2(t))dt, \quad (7.32a)$$

$$\text{subject to } \dot{x}(t) = -x(t) + u(t); \quad (7.32b)$$

$$x(0) = 1. \quad (7.32c)$$

Example 7.5.1 has been frequently considered in the literature as a benchmark for testing distinct computational methods. The admissible time varying OC  $u^*(t)$ , optimal state  $x^*(t)$ , and the optimal cost function  $J^*$  can be obtained via the necessary conditions of optimality in the following form:

$$u^*(t) = \frac{\sinh(\sqrt{2}(t-1))}{\sqrt{2} \cosh(\sqrt{2}) + \sinh(\sqrt{2})}, \quad (7.33)$$

$$x^*(t) = \cosh(\sqrt{2}t) - \frac{\sinh(\sqrt{2}t)(\sqrt{2} \tanh(\sqrt{2}) + 1)}{\tanh(\sqrt{2}) + \sqrt{2}}; \quad (7.34)$$

$$J^* = \frac{\sinh(\sqrt{2})}{2(\sinh(\sqrt{2}) + \sqrt{2} \cosh(\sqrt{2}))}. \quad (7.35)$$

To apply the GTM for solving the problem, let  $a = (a_1, \dots, a_n)^T$ ;  $b = (b_1, \dots, b_m)^T$ , and expand the state and the control variables by the Gegenbauer expansion series as follows:

$$x(t) \approx \sum_{k=0}^L a_k C_k^{(\alpha)}(t); \quad (7.36)$$

$$u(t) \approx \sum_{k=0}^M b_k C_k^{(\alpha)}(t). \quad (7.37)$$

## Chapter 7

---

Let  $v_{(k)}$  denotes the  $k$ -times multiple Hadamard product for any vector  $v$ , i.e.

$$v_{(k)} = \underbrace{v \circ v \circ \dots \circ v}_{k\text{-times}},$$

then the GTM presented in Section 7.3.1 transcribes the CTOCP into the following parameter NLP problem:

$$\text{minimize } J \approx \frac{1}{2} P_{N+1}^{(1)} \left( (\chi_{L,N+1}^{(\alpha)} a)_{(2)} + (\chi_{M,N+1}^{(\alpha)} b)_{(2)} \right), \quad (7.38a)$$

$$\text{subject to } \left[ \xi_L^{(\alpha)} + \wp_{L1}^{(\alpha)}, -\wp_{M1}^{(\alpha)} \right] (a, b)^T \approx \hat{e}_N, \quad (7.38b)$$

where “[.,.]” is the usual horizontal matrix concatenation notation. The NLP Problem (7.38) can be solved readily in the spectral space using MATLAB Optimization Toolbox. The state vector  $X = [x_0, x_1, \dots, x_N]^T$  and the control vector  $U = [u_0, u_1, \dots, u_N]^T$  can then be approximated at the GG points  $t_i \in S_N^{(\alpha)}$  through the relations:

$$X \approx \xi_L^{(\alpha)} a, \quad (7.39)$$

$$U \approx \xi_M^{(\alpha)} b, \quad (7.40)$$

where  $x_i = x(t_i); u_i = u(t_i) \forall i$ . Table 7.1 shows a comparison between the present GTM at  $S_5^{(0.5)}$  and some other direct optimization methods quoted from the literature. A comparison between the exact cost function value  $J^* \approx 0.192909298093169$  accurate to 15 decimal digits, and the approximate cost function value  $J$  obtained by the GTM with 5<sup>th</sup>-order Gegenbauer expansions of the state and the control variables, and a P-matrix constructed using  $M_P = 5$ , shows an agreement of 9 decimal figures. This is accomplished by recasting the CTOCP (7.32) into a 12-dimensional NLP problem using only 6 collocation points. The average CPU time in 100 runs taken by the GTM was found to be 0.367 seconds. Figure 7.1 shows the profiles of the exact control and state variables on  $[0, 1]$  versus the approximate control and the state variables. The dash-dotted lines represent the exact state and control solutions while the dotted lines represent the approximate state and control histories along the time horizon. Here the graph clearly demonstrates the high accuracy achieved by the GTM as we cannot distinguish between the exact and the approximate solutions by visual inspection. This demonstrates numerically the feasibility and the accuracy of the discrete optimal solutions obtained by the GTM. On the contrary, the same reported approximate cost function value was obtained by El-Gindy et al. (1995); Elnagar (1997); Razzaghi and Elnagar (1993); Vlassenbroeck and Dooren (1988) using larger numbers of spectral expansion terms, which lead to higher-dimensional NLP problems. For instance, the classical Chebyshev expansion method of Vlassenbroeck and Dooren

## Chapter 7

---

(1988) and the Shifted Legendre method of Razzaghi and Elnagar (1993) recast the CTOCP into a NLP problem of dimension 20 to obtain an approximate cost function value accurate to the same number of decimal digits. The El-Gindy et al. (1995) Chebyshev method leads to an 18 dimensional NLP problem while the cell averaging Chebyshev method of Elnagar (1997) produces a 16 dimensional NLP problem. We notice here that the Gegenbauer method of El-Hawary et al. (2003) produces an approximate cost function value which agrees with the exact cost function value only up to 7 decimal figures.

Example 7.5.1		
Methods	DIM	$J$
<b>Classical Chebyshev method (Vlassenbroeck and Dooren, 1988)</b>		
$m = 5, N = 8; K = 16$	12	0.1929094
$m = 7, N = 10; K = 20$	16	0.1929030
$m = 9, N = 15; K = 30$	20	0.1929092981
<b>Shifted Legendre method (Razzaghi and Elnagar, 1993)</b>		
$m = 5$	12	0.1929092980
$m = 9$	20	0.1929092981
<b>Chebyshev method (El-Gindy et al., 1995)</b>		
$m = 5; N = 5$	12	0.192881804
$m = 7; N = 7$	16	0.192906918
$m = 9; N = 9$	20	0.192909306
$m = 5; N = 11$	18	0.192909298
<b>Cell averaging Chebyshev method (Elnagar, 1997)</b>		
$m = 4$	10	0.19290924
$m = 5$	12	0.192909288
$m = 7$	16	0.1929092981
<b>Gegenbauer method (El-Hawary et al., 2003)</b>		
$M = N = 4; \alpha^* = 0.421$	11	0.192909281
<b>Present GTM</b>		
$L = M = M_P = 5$	12	0.1929092981277
$L = M = 5; M_P = 20$	12	0.1929092980933
<b>The optimal cost function value <math>J^* \approx 0.192909298093169</math></b>		

Table 7.1: The numerical results obtained by different methods for solving Example 7.5.1. DIM refers to the dimension of the NLP problem. The results of the present GTM are obtained at  $S_5^{(0.5)}$ .

Table 7.1 shows also that using a rectangular P-matrix with  $M_P = 20$  boosts the obtained accuracy of the GTM further to reach an agreement of 12 decimal figures while preserving the same dimension of the resulting NLP problem. The average CPU time taken by the GTM in this case was found to be 0.3045 seconds, which is shorter than the calculation time elapsed for  $M_P = 5$ . This suggests that the GTM can be implemented using the P-matrix with larger values of  $M_P$  than  $N$  to produce higher-order approximations in a short time without increasing the dimension of the NLP problem or adding any further constraints. Moreover, increasing the value of  $M_P$  support the rapid convergence of the GTM through the correct translation of the involved integral operations into more precise al-

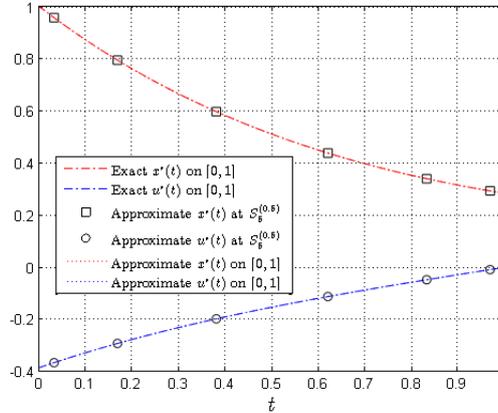


Figure 7.1: The numerical experiments of the GTM on Example 7.5.1. The figure shows the profiles of the exact control and state variables on  $[0, 1]$  together with the approximate optimal state and control variables obtained by the GTM. The results are reported at  $S_5^{(0.5)}$  using  $L = M = M_P = 5$ .

gebraic expressions. Hence the key element for these substantial results lies in the accurate approximations of the integral operators encountered during the discretization of the CTOCP without enlarging the dimension of the reduced NLP problem or escalating the required number of collocation points. These useful features of the GTM are vital to allow real-time decision making, and highlight the power of the proposed method. Moreover, the ability of the GTM to produce higher-order approximations quickly without affecting the dimensionality of the NLP problem or the number of included constraints are distinctive features of the GTM, which separate it from the rest of the available direct orthogonal collocation methods and direct pseudospectral methods in the literature. Notice here that although the GTM is carried out through collocation at the Legendre-Gauss points  $t_i \in S_5^{(0.5)}$ , the optimal P-matrix is not constructed using Legendre polynomial expansions. Instead, the developed optimal P-matrix takes on a pointwise approach by employing a distinct member of the Gegenbauer family of polynomials to optimally approximate the definite integral  $\int_0^{t_i} f(t)dt$ , for any smooth function  $f(t)$  and a certain collocation point  $t_i$ ; cf. (Elgindy and Smith-Miles, 2013b). For  $M_P > M_{\max}$ , the Gegenbauer quadrature associated with the optimal P-matrix becomes identical with the optimal Chebyshev and Legendre quadratures in the  $L^\infty$ -norm and  $L^2$ -norm, respectively.

The approximate optimal states and control variables obtained by the GTM

## Chapter 7

---

at  $S_5^{(0.5)}$  using  $L = M = 5; M_P = 20$  are given by:

$$x^*(t) = -\frac{3}{8000000000000000000} (-26666502823355615 + 36947921533280322t - 26591472235910150t^2 + 12003897953974440t^3 - 3813539900277835t^4 + 600671546504358t^5), \quad (7.41a)$$

$$u^*(t) = \frac{1}{8000000000000000000} (-30864900386971099 + 49117953959590954t - 30737975741949462t^2 + 15944062101269320t^3 - 4382058372586615t^4 + 923027310810486t^5), \quad (7.41b)$$

respectively. A sketch showing the profiles of the absolute errors  $E_x; E_u$  between the exact state and control variables and their approximations on the time horizon  $[0, 1]$  is shown in Figure 7.2. It can be clearly seen from the figure that the absolute errors  $E_x; E_u$  are small over the time horizon  $[0, 1]$ , and they reach their maximum values of  $6.14412 \times 10^{-6}; 7.34135 \times 10^{-6}$  at  $t = 0$ , respectively. Hence the optimal trajectories generated by the GTM are feasible and accurate.

### Example 7.5.2.

$$\text{minimize } J = \int_0^1 (x_1^2(t) + x_2^2(t) + 0.005u^2(t))dt, \quad (7.42a)$$

$$\text{subject to } \dot{x}_1(t) = x_2(t), \quad (7.42b)$$

$$\dot{x}_2(t) = -x_2(t) + u(t), \quad (7.42c)$$

$$x_1(0) = 0, \quad (7.42d)$$

$$x_2(0) = -1; \quad (7.42e)$$

$$x_2(t) - 8(t - 0.5)^2 + 0.5 \leq 0. \quad (7.42f)$$

Example 7.5.2 is a case model of a CTOCP with an inequality path constraint. The dynamics is a system of two linear differential equations provided with the initial conditions (7.42d) & (7.42e). The main goal is to find the OC  $u^*(t)$  which satisfies the dynamical system and Constraints (7.42d)-(7.42f) while minimizing the quadratic performance index (7.42a). Example 7.5.2 has no analytical solution, and the presence of the path inequality constraint (7.42f) adds more complexity over Example 7.5.1. Through the change of variable  $x = x_1$ , the CTOCP can be restated as follows:

$$\text{minimize } J = \int_0^1 (x^2(t) + \dot{x}^2(t) + 0.005u^2(t))dt, \quad (7.43a)$$

$$\text{subject to } \ddot{x}(t) + \dot{x}(t) - u(t) = 0, \quad (7.43b)$$

## Chapter 7

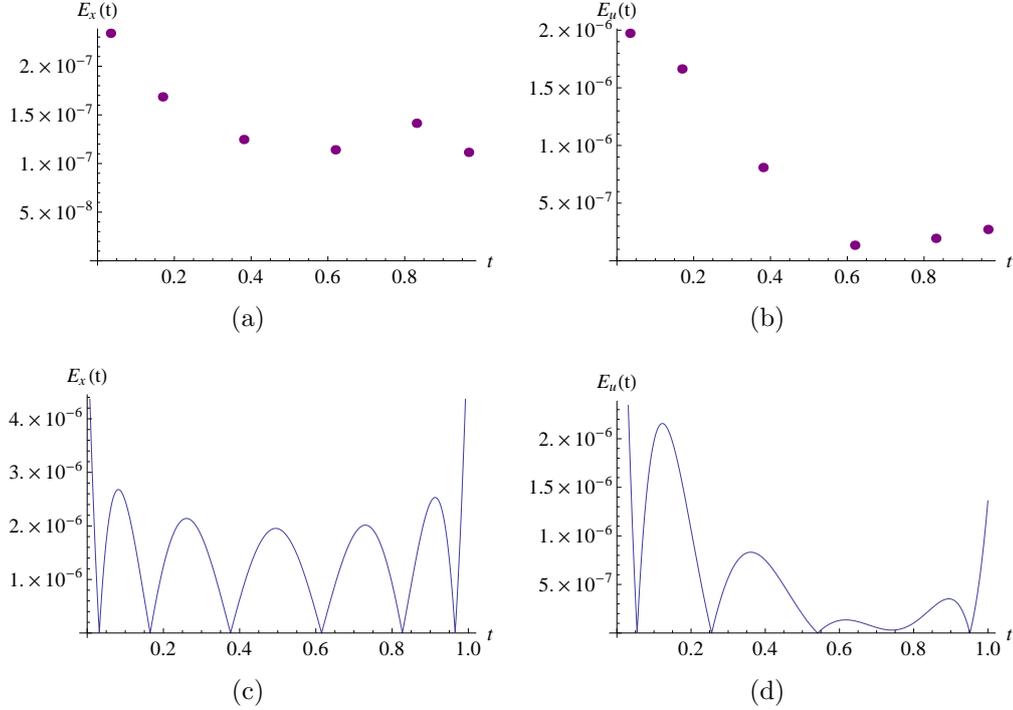


Figure 7.2: The sketch of the absolute errors  $E_x(t)$ ;  $E_u(t)$  obtained using Gegenbauer collocation at  $S_5^{(0.5)}$  with  $L = M = 5$ ;  $M_P = 20$ . Figures (a) & (b) show the values of the absolute errors  $E_x$ ;  $E_u$  at the GG collocation nodes while Figures (c) & (d) show the profiles of the absolute errors on the time interval  $[0, 1]$ . It can be clearly seen from the former two figures that the absolute errors are small at the GG points as expected. The latter two figures show that the absolute errors of the state and the control variables are also small over the whole time horizon with  $\max_{t \in [0,1]} E_x(t) \approx 6.14412 \times 10^{-6}$ ;  $\max_{t \in [0,1]} E_u(t) \approx 7.34135 \times 10^{-6}$ , respectively.

$$x(0) = 0, \quad (7.43c)$$

$$\dot{x}(0) = -1; \quad (7.43d)$$

$$\dot{x}(t) - 8(t - 0.5)^2 + 0.5 \leq 0. \quad (7.43e)$$

Following the GTM presented in Section 7.3.2, let  $\ddot{x}(t) = \mu(t)$ , for some unknown continuous function  $\mu(t)$ , and expand  $\mu(t)$  by a Gegenbauer expansion series in the following form:

$$\mu(t) \approx \sum_{k=0}^L a_k C_k^{(\alpha)}(t). \quad (7.44)$$

Also we expand the control variable  $u(t)$  by Equation (7.37). The GTM then

## Chapter 7

---

transcribes the CTOCP into the following finite-dimensional parameter NLP:

$$\text{minimize } J \approx P_{N+1}^{(1)} \left( (\theta_{L,2}^{(\alpha)} a - \sigma_{N+1})_{(2)} + (\theta_{L,1}^{(\alpha)} a - \hat{e}_{M_P})_{(2)} + 0.005(\chi_{M,N+1}^{(\alpha)} b)_{(2)} \right), \quad (7.45a)$$

$$\text{subject to } \left[ \xi_L^{(\alpha)} + \wp_{L,1}^{(\alpha)}, -\xi_M^{(\alpha)} \right] (a, b)^T \approx \hat{e}_N, \quad (7.45b)$$

$$\left[ \wp_{L,1}^{(\alpha)}, O \right] (a, b)^T \leq 8(\bar{t} - 0.5\hat{e}_N)_{(2)} + 0.5\hat{e}_N, \quad (7.45c)$$

where  $\sigma_{N+1} = (z_{N+1,0}, \dots, z_{N+1,M_P})^T$ ,  $O \in \mathbb{R}^{(N+1) \times (M+1)}$  is the  $(N+1) \times (M+1)$  zero matrix;  $\bar{t} = (t_0, \dots, t_N)^T$  is the solution nodes vector. The NLP problem (7.45) can be solved in the spectral space, and the state vector  $X$ , its derivative  $\dot{X}$  together with the control vector  $U$  at the GG collocation points  $\{t_i\}_{i=0}^N$  can be approximated by the following relations:

$$X \approx \wp_{L,2}^{(\alpha)} a - \bar{t}, \quad (7.46)$$

$$\dot{X} \approx \wp_{L,1}^{(\alpha)} a - \hat{e}_N; \quad (7.47)$$

$$U \approx \xi_M^{(\alpha)} b. \quad (7.48)$$

The state vectors  $X_r = (x_r(t_0), \dots, x_r(t_N))^T$ ;  $r = 1, 2$  of the original CTOCP (7.42) are then obtained by setting

$$X_1 = X, \quad (7.49a)$$

$$X_2 = \dot{X}, \quad (7.49b)$$

respectively. Moreover, the optimal state variables  $x_1^*(t); x_2^*(t)$  can be directly recovered over the entire time interval  $[0, 1]$  by successive integrations through Equations (7.21) as follows:

$$x_1^*(t) = \int_0^t \int_0^{\tau_2} \mu(\tau_1) d\tau_1 d\tau_2 - t = \sum_{k=0}^L a_k \int_0^t \int_0^{\tau_2} C_k^{(\alpha)}(\tau_1) d\tau_1 d\tau_2 - t; \quad (7.50a)$$

$$x_2^*(t) = \int_0^t \mu(\tau_1) d\tau_1 - 1 = \sum_{k=0}^L a_k \int_0^t C_k^{(\alpha)}(\tau_1) d\tau_1 - 1. \quad (7.50b)$$

The OC variable  $u^*(t)$  can be evaluated directly via Equation (7.37). Figure 7.3 shows the approximate optimal state and control variables profiles on  $[0, 1]$  obtained by the GTM at  $S_8^{(0.2)}$  using  $L = M = 6$ . It can be clearly seen from the figure that the  $x_2$ -trajectory starts off away from the boundary constraint  $r(t) = 8(t - 0.5)^2 - 0.5$ , and then approaches it very quickly. The trajectory then

## Chapter 7

---

matches the constraint boundary along the mid time interval  $[0, 1]$  for about 0.4 seconds. Eventually, the trajectory leaves the constraint boundary, and becomes stable at about zero value.

Table 7.2 shows a comparison between the present GTM using the P-matrix with  $M_P = 20$  and some other numerical methods quoted from the literature. Here we notice from the table that the GTM converges rapidly to the approximate cost function value  $J^* \approx 0.17$  for increasing numbers of Gegenbauer expansion terms and collocation nodes while maintaining significantly small-scale NLP problems. Moreover, the reported optimal cost function value  $J^*$  of the GTM is lower than those obtained by Elnagar (1997); Jaddu (2002); Marzban and Razzaghi (2003, 2010); Mashayekhi et al. (2012); Vlassenbroeck (1988); Yen and Nagurka (1991) as evident from Table 7.2 with a high degree of constraint satisfaction. Furthermore, for the best results obtained in all methods, the dimension of the resulting NLP problem produced by the GTM is about half that of Jaddu (2002); Vlassenbroeck (1988); Yen and Nagurka (1991), one-third that of Marzban and Razzaghi (2003), two-fifths that of Mashayekhi et al. (2012); 6% that of Marzban and Razzaghi (2010). The table shows also that the 6<sup>th</sup>-order Gegenbauer expansions of the states and the control is sufficient to retain the first 4 decimal figures of the approximate optimal cost function value  $J^*$ . Notice here that the GTM discretizes the CTOCP using only 9 collocation points, and produces a significantly small-scale NLP problem of dimension 14. The reported average CPU time taken by the GTM was found to be 0.2978 seconds. Hence the GTM converges quickly using relatively small number of collocation points and expansion terms.

The approximate optimal states and the control variables obtained by the

## Chapter 7

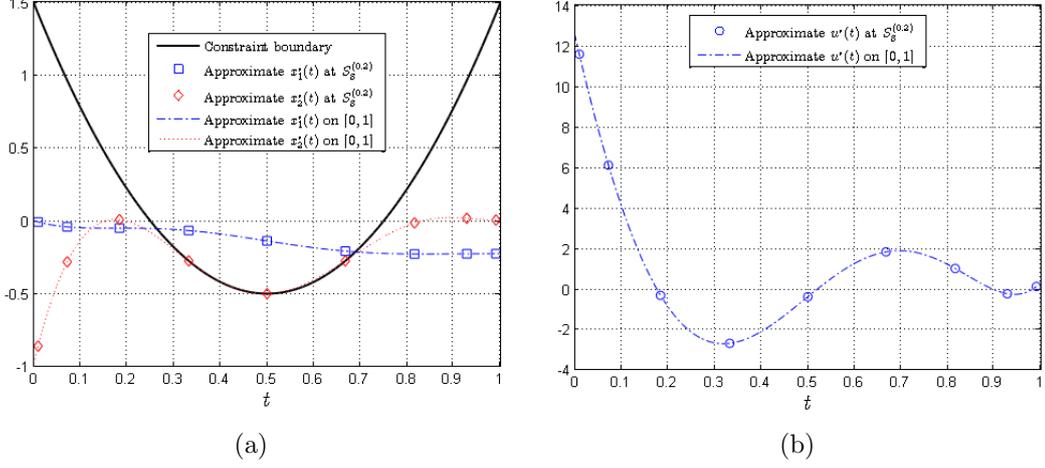


Figure 7.3: The numerical experiments of the GTM on Example 7.5.2. Figures (a) and (b) show the graphs of the approximate optimal state and control variables on  $[0, 1]$ . The results of the GTM are obtained at  $S_g^{(0.2)}$  using  $L = M = 6$ . The solid, dotted, and dash-dotted lines are generated using 100 nodes.

GTM through collocation at  $S_g^{(0.3)}$  with  $L = M = 8$  are given by:

$$x_1^*(t) = \frac{1}{107251200000000000000000000000000} (107251200000000000000000000000000 - 749670613091978780125875t + 2317581710580779575023000t^2 - 3100276209501909528937000t^3 + 2203431341043705119379900t^4 - 1287743709209664232402450t^5 + 288863004777584209303800t^6 + 858523307501187124024650t^7 - 793212720489943685375150t^8 + 179753341502203543032861t^9), \quad (7.51a)$$

$$x_2^*(t) = \frac{1}{107251200000000000000000000000000} (-107251200000000000000000000000000 + 149934122618395756025175t - 695274513174233872506900t^2 + 1240110483800763811574800t^3 - 1101715670521852559689950t^4 + 772646225525798539441470t^5 - 202204103344308946512660t^6 - 686818646000949699219720t^7 + 713891448440949316837635t^8 - 179753341502203543032861t^9); \quad (7.51b)$$

$$u^*(t) = \frac{1}{127680000000000000000000000000000} (1657218378200811721595 - 14765806523585040468720t + 35932057440845058264040t^2 - 36921021079930681285200t^3 + 29003619222672898873150t^4 + 5724452685786000717600t^5 - 78050553605280800587000t^6 + 77912337906040804081440t^7 - 20390222164737752460777t^8). \quad (7.51c)$$

## Chapter 7

Example 7.5.2		
Methods	DIM	$J$
<b>Classical Chebyshev method (Vlassenbroeck, 1988)</b>		
$m = 11, K = 24$	36	0.17784
$m = 12, K = 26$	39	0.17358
$m = 13, K = 28$	42	0.17185
<b>Fourier-based state parameterization (Yen and Nagurka, 1991)</b>		
$K = 5$	28	0.17115
$K = 7$	36	0.17069
$K = 9$	44	0.17013
<b>Cell averaging Chebyshev method (Elnagar, 1997)</b>		
$m = 5$	12	0.17350546
$m = 7$	16	0.17185501
$m = 9$	20	0.17184981
<b>Chebyshev method (Jaddu, 2002)</b>		
$N = 13$	42	0.17078488
<b>Hybrid block-pulse and Legendre method (Marzban and Razzaghi, 2003)</b>		
$N = 4, M = 3; w = 15$	60	0.17013645
$N = 4, M = 4; w = 15$	80	0.17013640
<b>Rationalized Haar method (Marzban and Razzaghi, 2010)</b>		
$K = 16; w = 100$	49	0.171973
$K = 32; w = 100$	97	0.170185
$K = 64; w = 100$	193	0.170115
$K = 128; w = 100$	385	0.170103
<b>Hybrid block-pulse functions and Bernoulli polynomials method (Mashayekhi et al., 2012)</b>		
$N = 4; M = 2$	36	0.1700316
$N = 4; M = 3$	48	0.1700305
$N = 4; M = 4$	60	0.1700301
<b>Present GTM</b>		
$\alpha = 0; N = L = M = 5$	12	0.170593
$\alpha = 0.2; N = 8; L = M = 6$	14	0.17008
$\alpha = 0.3; N = L = M = 8$	18	0.170052
$\alpha = 0.2; N = 9; L = M = 8$	18	0.16998
$\alpha = 0.9; N = 11; L = M = 10$	22	0.170039
$\alpha = 0.9; N = L = M = 11$	24	0.17

Table 7.2: The approximate cost function value  $J$  of Example 7.5.2 obtained by different methods. The results of the present GTM are obtained at the GG collocation set  $S_N^{(\alpha)}$ , for the shown values of  $\alpha; N$  using different values of  $L; M$ .

The average elapsed CPU time in this case was found to be 0.622 seconds. Evaluating the optimal state variables  $x_1(t); x_2(t)$  at  $t = 0$ , respectively, using exact arithmetic in MATHEMATICA 8 software Version 8.0.4.0 show that our optimal state solutions perfectly match the initial conditions with zero error. To check the satisfaction of the dynamical system equations (7.42b) & (7.42c), and the inequality constraint (7.42f), let

$$\mathcal{E}_1(t) = \dot{x}_1(t) - x_2(t), \quad (7.52a)$$

$$\mathcal{E}_2(t) = \dot{x}_2(t) + x_2(t) - u(t); \quad (7.52b)$$

$$\mathcal{E}_3(t) = x_2(t) - 8(t - 0.5)^2 + 0.5. \quad (7.52c)$$

Then we can clearly verify that  $\mathcal{E}_1(t) = 0 \forall t \in [0, 1]$ , and Equation (7.42b) is perfectly satisfied over the whole time interval. The sketches of the error function  $\mathcal{E}_2(t)$  and the inequality constraint function  $\mathcal{E}_3(t)$  are shown in Figure 7.4. Figure 7.4(a) shows the magnitude of the error function  $\mathcal{E}_2(t)$  at the collocation nodes  $\{t_i\}_{i=0}^8$ , where the optimal solutions achieve highly constraint satisfaction as expected. Figure 7.4(b) shows the profile of the error function over the time

## Chapter 7

---

horizon  $[0, 1]$ , where we can clearly see that the magnitude of the error is small, and reaches its maximum value of approximately  $2.479 \times 10^{-4}$  at  $t = 0$ . Figure 7.4(c) confirms the satisfaction of the inequality constraint (7.42f) on the entire time horizon  $[0, 1]$ , where the optimal  $x_2$ -trajectory never violate the boundary constraint  $r(t)$ . Hence the optimal trajectories obtained by the GTM are feasible. Moreover, the GTM converges to a lower cost function value than the other traditional methods of Elnagar (1997); Jaddu (2002); Marzban and Razzaghi (2003, 2010); Mashayekhi et al. (2012); Vlassenbroeck (1988); Yen and Nagurka (1991). These results show that the GTM offers many useful advantages over the available discretization methods, and the power of the proposed GTM is conspicuous in the achievement of higher-order approximations using relatively small number of Gegenbauer expansion terms and collocation nodes. The simplicity, rapid convergence, and the precise approximations of the GTM make its application for the solution of a wide variety of CTOCPs quite attractive and beneficial.

**Example 7.5.3.** Consider the problem of finding the OC  $u^*(t)$  which minimizes the performance index (7.42a) subject to Constraints (7.42b)-(7.42e), and the following state inequality constraint:

$$x_1(t) - 8(t - 0.5)^2 + 0.5 \leq 0. \quad (7.53)$$

Here the GTM produces the following NLP problem:

$$\text{minimize } J \approx P_{N+1}^{(1)} \left( (\theta_{L,2}^{(\alpha)} a - \sigma_{N+1})_{(2)} + (\theta_{L,1}^{(\alpha)} a - \hat{e}_{M_P})_{(2)} + 0.005(\chi_{M,N+1}^{(\alpha)} b)_{(2)} \right), \quad (7.54a)$$

$$\text{subject to } \left[ \xi_L^{(\alpha)} + \varphi_{L,1}^{(\alpha)}, -\xi_M^{(\alpha)} \right] (a, b)^T \approx \hat{e}_N; \quad (7.54b)$$

$$\left[ \varphi_{L,2}^{(\alpha)}, O \right] (a, b)^T \leq 8(\bar{t} - 0.5\hat{e}_N)_{(2)} + \bar{t} - 0.5\hat{e}_N. \quad (7.54c)$$

The control vector  $U$  and the state vectors  $X_1; X_2$  are approximated through Equations (7.48) & (7.49) as described before in Example 7.5.2. The GTM converges to the approximate optimal cost function value  $J^* \approx 0.71865$ . The average elapsed CPU time taken by the GTM through collocation at  $S_8^{(-0.4)}$  with  $L = M = 12$  was found to be 0.65 seconds. Figure 7.5 shows the corresponding state and control profiles on the time horizon  $[0, 1]$ . It can be seen from the figure that the  $x_1$ -trajectory decreases linearly during the first half of the time domain till it touches the constraint boundary along the neighborhood of  $t = 0.5$ . The trajectory then becomes identical with the tangent of the constraint boundary at  $t = 0.5$ , and remains constant about the value  $-0.5$ . Table 7.3 shows a comparison between the present GTM and the methods of Elnagar (1997); Foroozandeh and Shamsi (2012); Vlassenbroeck (1988). The numerical

## Chapter 7

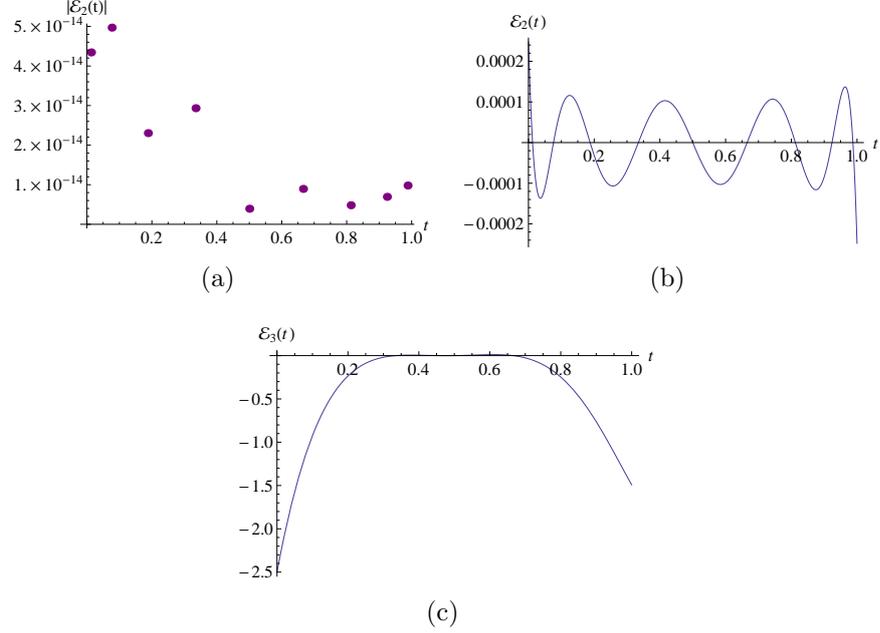


Figure 7.4: The plots of the error function  $\mathcal{E}_2(t)$  and the inequality constraint function  $\mathcal{E}_3(t)$  produced by the present GTM through Gegenbauer collocation at  $S_8^{(0.3)}$  with  $L = M = 8$ . Figure (a) shows the magnitude of the error function  $\mathcal{E}_2(t)$  at the GG collocation nodes  $t_i \in S_8^{(0.3)}$ . Figure (b) shows the propagation of the error function  $\mathcal{E}_2(t)$  on the time interval  $[0, 1]$ , where it can be clearly seen that the error function is an oscillatory function of small magnitude over the whole time horizon with  $\max_{t \in [0, 1]} |\mathcal{E}_2(t)| \approx 2.479 \times 10^{-4}$ . Figure (c) shows the profile of the nonnegative inequality constraint function  $\mathcal{E}_3(t)$  on the time interval  $[0, 1]$ , where it can be verified that the optimal  $x_2$ -trajectory never cross the boundary constraint  $r(t)$ .

results suggest that the GTM performs better than the conventional discretization methods, and produces a significantly smaller-scale NLP problem for many suitable choices of the parameters  $\alpha, N, L; M$ .

### Example 7.5.4.

$$\text{minimize } J = \int_0^1 (x_1^4(t) + x_2^4(t) + u^4(t))dt, \quad (7.55a)$$

$$\text{subject to } \dot{x}_1(t) = x_2(t), \quad (7.55b)$$

$$\dot{x}_2(t) = -0.032x_2(t) - 0.16x_1(t) + 1.6u(t); \quad (7.55c)$$

$$(\dot{x}_1(0), x_1(1)) = (1, 0). \quad (7.55d)$$

## Chapter 7

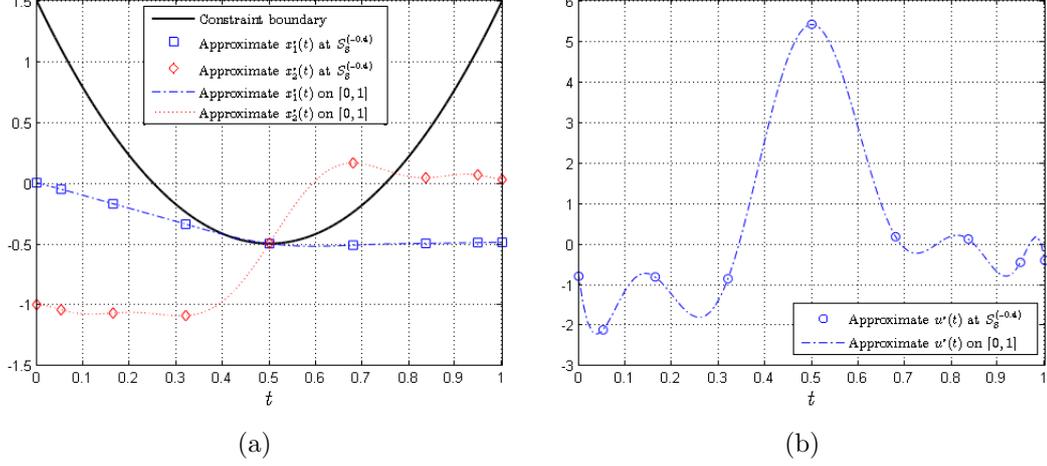


Figure 7.5: The numerical experiments of the GTM on Example 7.5.3. Figures (a) and (b) show the graphs of the approximate state variables and the control variable on  $[0, 1]$ . The results are obtained at  $S_8^{(-0.4)}$  using  $L = M = 12$ . The solid, dotted, and dash-dotted lines are generated using 100 nodes.

Example 7.5.4 is a further challenging CTOCP which has no analytical solution. Here the Lagrangian function is a quartic function in the states and the control. To solve the CTOCP numerically using the GTM, we introduce the change of variable  $x = x_1$ , and restate the problem as follows:

$$\text{minimize } J = \int_0^1 (x^4(t) + \dot{x}^4(t) + u^4(t))dt, \quad (7.56a)$$

$$\text{subject to } \ddot{x}(t) + 0.032\dot{x}(t) + 0.16x(t) - 1.6u(t) = 0; \quad (7.56b)$$

$$(\dot{x}(0), x(1)) = (1, 0). \quad (7.56c)$$

Problem (7.56) can be transformed by the GTM into the following constrained NLP problem:

$$\begin{aligned} \text{minimize } J \approx P_{N+1}^{(1)} & \left( ((\theta_{L_2}^{(\alpha)} - \hat{e}_{M_P} \bar{\varphi}_{L_2}^{(\alpha)})a + \sigma_{N+1} - \hat{e}_{M_P})_{(4)} + (\theta_{L_1}^{(\alpha)} a + \hat{e}_{M_P})_{(4)} \right. \\ & \left. + (\chi_{M, N+1}^{(\alpha)} b)_{(4)} \right), \end{aligned} \quad (7.57a)$$

$$\text{subject to } \left[ \xi_L^{(\alpha)} + 0.16(\varphi_{L_2}^{(\alpha)} - \hat{e}_N \bar{\varphi}_{L_2}^{(\alpha)}) + 0.032\varphi_{L_1}^{(\alpha)}, -1.6\xi_M^{(\alpha)} \right] (a, b)^T \approx 0.128\hat{e}_N - 0.16\bar{t}, \quad (7.57b)$$

where

$$\bar{\varphi}_{Lq}^{(\alpha)} = (P_{N+1}^{(q)} \bar{\zeta}_{L, N+1, 0}^{(\alpha)}, \dots, P_{N+1}^{(q)} \bar{\zeta}_{L, N+1, L}^{(\alpha)}) \approx (I_{q, 0}^{(\alpha)}(t_{N+1}), \dots, I_{q, L}^{(\alpha)}(t_{N+1})).$$

## Chapter 7

Example 7.5.3		
Methods	DIM	$J$
<b>Classical Chebyshev method (Vlassenbroeck, 1988)</b>		
$m = 5; K = 12$	18	0.76600000
$m = 9; K = 20$	30	0.74830000
$m = 11; K = 24$	36	0.74522000
$m = 13; K = 28$	42	0.74096000
<b>Cell averaging Chebyshev method (Elnagar, 1997)</b>		
$m = 5$	12	0.74032810
$m = 7$	16	0.74088140
$m = 9$	20	0.74096103
<b>Interpolating scaling functions method (Foroozandeh and Shamsi, 2012)</b>		
$n = 1; r = 4$	24	0.74605803
$n = 3; r = 4$	96	0.73861271
$n = 3; r = 5$	120	0.73740941
$n = 4; r = 5$	240	0.73744874
$n = 5; r = 5$	480	0.73743852
<b>Present GTM</b>		
$\alpha = 1; N = L = M = 4$	10	0.731002
$\alpha = 1; N = L = M = 6$	14	0.728105
$\alpha = 0.9; N = L = M = 8$	18	0.723743
$\alpha = 0, N = 8; L = M = 10$	22	0.720437
$\alpha = 0.8, N = 8; L = M = 11$	24	0.719507
$\alpha = -0.4, N = 8; L = M = 12$	26	0.718658
$\alpha = 0, N = 10; L = M = 13$	28	0.718653
$\alpha = -0.4, N = 14; L = M = 15$	32	0.718654
$\alpha = 0.8; N = L = M = 16$	34	0.71865

Table 7.3: The approximate cost function of Example 7.5.3 obtained by different methods.

The constrained NLP problem (7.57) can be easily solved in the spectral space using MATLAB Optimization Toolbox. The values of the states  $x_1(t); x_2(t)$ , and the control variable  $u(t)$  can be obtained at the GG solution points  $t_i \in S_N^{(\alpha)}$  through the following equations:

$$X_1 = X \approx (\wp_{L2}^{(\alpha)} - \hat{e}_N \bar{\wp}_{L2}^{(\alpha)})a + \bar{t} - \hat{e}_N, \quad (7.58a)$$

$$X_2 = \dot{X} \approx \wp_{L1}^{(\alpha)}a + \hat{e}_N; \quad (7.58b)$$

$$U \approx \xi_M^{(\alpha)}b. \quad (7.58c)$$

Moreover, the optimal paths of the states and the control variables can be obtained at any intermediate point in the time horizon  $[0, 1]$  using the following equations:

$$x_1^*(t) \approx \sum_{k=0}^L a_k (I_{2,k}^{(\alpha)}(t) - P_{N+1}^{(2)} \bar{\zeta}_{L,N+1,k}^{(\alpha)}) + t - 1, \quad (7.59a)$$

$$x_2^*(t) \approx \sum_{k=0}^L a_k I_{1,k}^{(\alpha)}(t) + 1, \quad (7.59b)$$

## Chapter 7

and Equation (7.37) without invoking any interpolation method. Comparisons between the present GTM and the piecewise polynomial control parameterization method of Sirisena and Tan (1974), Legendre pseudospectral method of Elnagar et al. (1995), Chebyshev pseudospectral method of Fahroo and Ross (2002), and the Chebyshev orthogonal collocation method of Ma et al. (2011) are shown in Table 7.4. The first column shows the optimal cost function value obtained by Sirisena and Tan (1974). The next three columns show the results of the Elnagar et al. (1995); Fahroo and Ross (2002), and Ma et al. (2011) methods listed in the form  $N/J$ , where  $(N + 1)$  is the number of collocation points, or the number of spectral expansion terms in the parameterization of the states and the control. The last column shows our numerical results listed in the form  $N/(L = M)/\alpha/J$ . To present the strength of the attractive rectangular form of the P-matrix, our results are conveniently divided further into two sub-columns: (i) The first sub-column shows the results of the GTM using a square P-matrix with  $M_P = N$ , for some values of  $N$  in the range  $[2, 256]$ . The second sub-column shows the results of the GTM using a rectangular P-matrix with the number of its columns ( $M_P + 1$ ) fixed, while the number  $N$  is allowed to increase; in particular, we set  $M_P = 20$ .

Example 7.5.4					
Piecewise polynomial method (Sirisena and Tan, 1974) $J$	Legendre pseudospectral method (Elnagar et al., 1995) $N/J$	Chebyshev pseudospectral method (Fahroo and Ross, 2002) $N/J$	Chebyshev orthogonal collocation method (Ma et al., 2011) $N/J$	Present GTM $N/(L = M)/\alpha/J$	
				$M_P = N$	$M_P = 20$
1.1975	-	-	10/1.1974	2/2/1/0.30897	2/2/1/0.298102
	64/1.197180	64/1.197397	64/1.197396	3/3/1/0.298565	3/3/-0.2/0.29809
	128/1.1971808	128/1.1973969	128/1.1973965	4/4/1/0.297796	4/4/1/0.298081
	256/1.19718137	256/1.19739754	256/1.19739648	5/5/-0.2/0.298079	5/5/0/0.298079
				6/6/1/0.298076	6/6/1/0.298078
				10/10/0.9/0.298078	10/10/0.8/0.298078
				64/10/0.5/0.2982	64/10/0.5/0.298078
				64/20/0.5/0.298263	64/20/0.3/0.298078
				128/10/0.5/0.298078	128/10/0.6/0.298078
				128/20/0.5/0.298099	128/20/-0.3/0.298078
				256/10/0.5/0.298081	256/10/0.3/0.298078
				256/20/0.5/0.298078	256/20/-0.3/0.298078

Table 7.4: Comparisons between the present GTM using the P-matrix and the methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011); Sirisena and Tan (1974). The reported results of the Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011) methods were obtained using SNOPT (Gill et al., 2005), and are exactly as quoted from Ref. (Ma et al., 2011).

Table 7.4 shows that the present GTM converges rapidly to local optimal solutions associated with a much lower value of the performance index  $J$  than the reported values in (Elnagar et al., 1995; Fahroo and Ross, 2002; Ma et al., 2011; Sirisena and Tan, 1974) for both cases of  $M_P = N; 20$ . This is due to the efficient discretization of the CTOCP (7.55) using the GTM into a signifi-

## Chapter 7

---

N	CPU time (Seconds)				
	Legendre pseudospectral method (Elnagar et al., 1995)	Chebyshev pseudospectral method (Fahroo and Ross, 2002)	Chebyshev orthogonal collocation method (Ma et al., 2011)	Present GTM	
				$M_P = 16$	$M_P = 20$
64	144.63	164.88	29.66	0.366	0.454
128	523.78	201.09	157.75	3.519	2.593
256	1144.49	1087.55	615.14	4.808	14.024

Table 7.5: The CPU time of the present method using the P-matrix with  $M_P = 16, 20$ ;  $L = M = 10$ , versus the methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011). The results of the present method for  $N = 64, 128; 256$  were obtained by collocations at the GG sets  $S_{64}^{(-0.1)}, S_{128}^{(0.5)}; S_{256}^{(0.3)}$  using  $M_P = 16$  and by collocations at the GG sets  $S_{64}^{(0.5)}, S_{128}^{(0.6)}; S_{256}^{(0.3)}$  using  $M_P = 20$ . The reported CPU times of the Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011) methods are exactly as quoted from Ref. (Ma et al., 2011).

cantly small-scale NLP problem, where a standard optimization solver such as the “fmincon” MATLAB optimization solver was able to determine better local optimal solutions than the other discretization methods. Moreover, the proposed numerical scheme remains stable in both cases  $M_P = N; 20$ , for large-scale number of collocation nodes. It can be noticed from the table that for as small as 3 collocation nodes, the present method produces plausible approximate solutions using 2nd-order Gegenbauer polynomial expansions, especially for  $M_P = 20$ . For  $M_P = N$ , the GTM converges to the approximate minimum cost function value  $J \approx 0.298078$  using 11 collocation nodes, and 10<sup>th</sup>-order Gegenbauer polynomial expansions. On the other hand, the GTM carried out using a P-matrix with  $M_P = 20$  shows a faster convergence rate as evidently seen from the table, where it converges to the same optimal cost function value using 7 collocation nodes, and 6<sup>th</sup>-order Gegenbauer polynomial expansions. The corresponding optimal

## Chapter 7

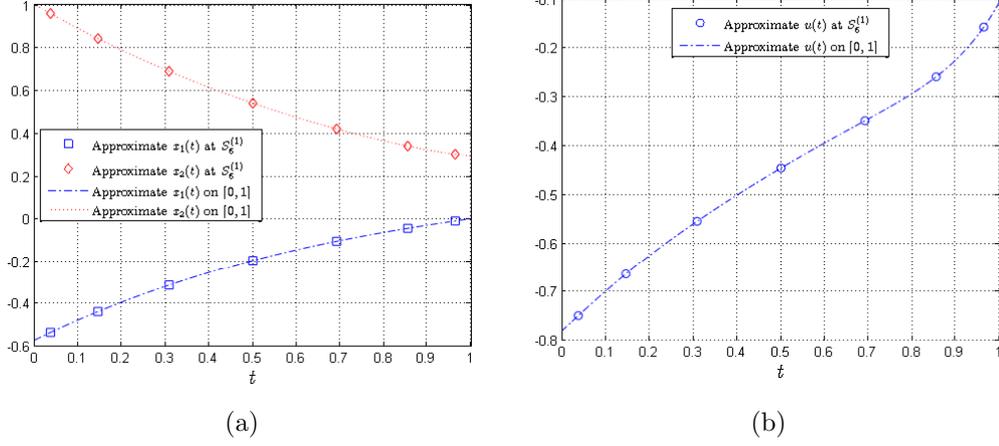


Figure 7.6: The numerical experiments of the GTM on Example 7.5.4. Figures (a) and (b) show the profiles of the states and the control variables on  $[0, 1]$ , respectively. The results are reported at  $S_6^{(1)}$  using  $M_P = 20$ ;  $L = M = 6$ .

states and control variables obtained by the GTM in this case are:

$$x_1^*(t) = \frac{1}{441 \times 10^{17}} \left( -25208106218129031404 + 441 \times 10^{17}t - 26235554693334864255t^2 + 8890691198465937735t^3 + 238515955440244860t^4 - 8673372878502419304t^5 + 15234497607142308768t^6 - 12308404906746959920t^7 + 3961733935664783520t^8 \right), \quad (7.60a)$$

$$x_2^*(t) = \frac{1}{441 \times 10^{17}} \left( 441 \times 10^{17} - 52471109386669728510t + 26672073595397813205t^2 + 954063821760979440t^3 - 43366864392512096520t^4 + 91406985642853852608t^5 - 86158834347228719440t^6 + 31693871485318268160t^7 \right), \quad (7.60b)$$

$$u^*(t) = \frac{1}{15 \times 10^{16}} \left( -117118600068747207 + 124777787017762440t - 391301269799916t^2 - 368807534085601010t^3 + 976214465291699328t^4 - 1104158142660680080t^5 + 473786969000696640t^6 \right), \quad (7.60c)$$

and a sketch showing the paths of the optimal solutions on the time horizon  $[0, 1]$  is shown in Figure 7.6. The state variable  $x_1(t)$ , and its derivative  $\dot{x}_1(t)$  utterly match the boundary conditions with zero error. To check the satisfaction of the dynamical system, let

$$\mathcal{E}_1(t) = \dot{x}_1(t) - x_2(t); \quad (7.61a)$$

$$\mathcal{E}_2(t) = \dot{x}_2(t) + 0.032x_2(t) + 0.16x_1(t) - 1.6u(t). \quad (7.61b)$$

## Chapter 7

---

Then we can clearly verify that  $\mathcal{E}_1(t) = 0 \forall t \in [0, 1]$ , and Equation (7.55b) is satisfied over the entire time interval. A sketch of the error function  $\mathcal{E}_2(t)$  is shown in Figure 7.7. Figure 7.7(a) shows the magnitude of the error function  $\mathcal{E}_2(t)$  at the collocation nodes  $\{t_i\}_{i=0}^6$ , where the optimal solutions achieve highly constraint satisfaction as expected. Figure 7.7(b) shows the profile of the error function over the time horizon  $[0, 1]$ , where we can clearly see that the magnitude of the error is small, and reaches its maximum value of  $2.1 \times 10^{-5}$  at  $t = 1$ . Hence the optimal trajectories obtained by the GTM are feasible, and produce a significantly lower cost function value than the other traditional methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011); Sirisena and Tan (1974).

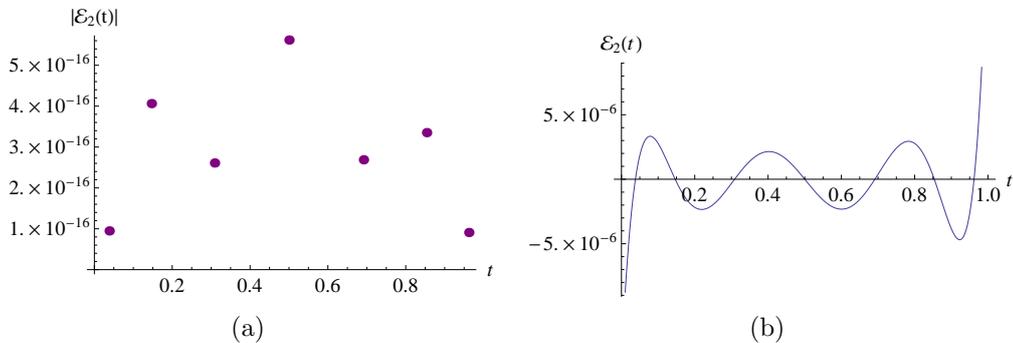


Figure 7.7: The sketch of the error function  $\mathcal{E}_2(t)$  built using Gegenbauer collocation at  $S_6^{(1)}$  with  $M_P = 20; L = M = 6$ . Figure (a) shows the magnitude of the error at the GG collocation nodes. Figure (b) shows the profile of the error function  $\mathcal{E}_2(t)$  on the time interval  $[0, 1]$ , where it can be clearly seen that the error function is an oscillatory function of small magnitude over the whole time horizon with  $\max_{t \in [0, 1]} |\mathcal{E}_2(t)| \approx 2.1 \times 10^{-5}$ .

Using the square P-matrix with  $M_P > M_{\max} = 20$ , the GTM collocates the OC problem at the Legendre-Gauss points  $S_N^{(0.5)}$ . The states and the control variables are expanded by Legendre expansion series, and the P-matrix is constructed by Algorithm 2.2 in (Elgindy and Smith-Miles, 2013b) via Legendre polynomial expansions to maintain robustness. On the other hand, using the rectangular P-matrix with  $M_P = 20$  yields a very stable numerical scheme for an increasing number of collocation points as clearly seen from Table 7.4, where the effect of the round-off error becomes very limited. Therefore the rectangular form of the P-matrix is extremely useful in the sense that: “For a small/large-scale number of collocation points, the robust GTM can converge rapidly to the local optimal solutions by choosing a suitable value of  $M_P$  without affecting the dimensionality of the resulting NLP problem.” Here it is noteworthy to mention that many of the

## Chapter 7

---

reported results in Table 7.4 maybe achieved using smaller values of  $M_P$  as well. For instance, our numerical experiments report that the same value of the approximate optimal cost function  $J \approx 0.298078$  is achievable for  $N = 64, 128; 256$  using  $L = M = 10; M_P = 16$  through collocations at  $S_{64}^{(-0.1)}, S_{128}^{(0.5)}, S_{256}^{(0.3)}$ . Table 7.5 shows a comparison between the elapsed CPU time of the present method versus that of the methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011). The reported CPU times suggest that the present GTM endowed with the optimal P-matrix can significantly reduce the required calculation time typically required by standard direct optimization methods. We notice here that the GTM carried out using  $M_P = 20$  required longer execution times compared to the lower value of  $M_P = 16$ , except for  $N = 128$ , where it converged faster.

The above experimental analysis shows that the GTM, as inherited from the application of the P-matrix, has the ability to produce higher-order approximations and faster convergence rates without the requirement of increasing the number of collocation nodes and the number of spectral expansion terms. In contrast, the traditional methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011) can only produce higher-order approximations by increasing the order of the Chebyshev and Legendre polynomials expansions in the approximations of the states and the control, in addition to the number of collocation nodes, which must be equal to that of the spectral expansion terms. For instance, collocating the CTOCP (7.55) using  $N = 256$  via the methods of Elnagar et al. (1995); Fahroo and Ross (2002); Ma et al. (2011) leads to an enormous NLP of dimension 771, which requires an efficient NLP solver tailored particularly for large-scale optimization problems. Moreover, the order of the traditional square spectral matrices employed in the approximations in (Elnagar et al., 1995; Fahroo and Ross, 2002) becomes large,  $O(257)$ , affecting the storage requirements, and intensifying the matrix algebra computations necessary. In contrast, the present method breaks the long standing bond between the number of collocation nodes, the number of spectral expansion terms, and the order of the employed spectral matrix, as the user has the freedom to choose any suitable values of  $L, M; M_P$ , for the same number of collocation nodes ( $N + 1$ ). For instance, for  $N = 256; M_P = 16$ , the order of the P-matrix becomes  $O(257 \times 17)$ , which is approximately 0.066 smaller than the number of storage requirements typically used by a traditional spectral matrix.

Another essential element in reducing the dimension of the resulting NLP, in addition to the rectangular form of the P-matrix integrated with the present GTM, lies in the substitution of the highest-order derivative occurring in the OC problem with some unknown function  $\mu(t)$ . This key idea does not only provide the luxury of working in a full integration environment, but rather significantly limits the number of optimization variables in the reduced NLP. To illustrate further, consider the very same CTOCP (7.55) discussed thoroughly above. A

## Chapter 7

---

typical direct orthogonal/pseudospectral method attempts to parameterize both the states  $x_1(t); x_2(t)$ , in addition to the control variable  $u(t)$ . For instance, if we assume equal number of states expansion terms  $L$ , then the states and the controls are typically expanded as follows:

$$x_r(t) \approx \sum_{k=0}^L a_{r,k} \phi_k^{(\alpha)}(t), \quad r = 1, 2, \quad (7.62a)$$

$$u(t) \approx \sum_{k=0}^M b_k \phi_k^{(\alpha)}(t), \quad (7.62b)$$

for some prescribed set of global orthogonal basis polynomials  $\{\phi_k^{(\alpha)}(t)\}$ , frequently chosen as the Chebyshev or Legendre polynomials. In this case, it is not hard to realize that the reduced NLP problem is of dimension  $(2L+M+3)$  opposed to only  $(L+M+2)$  as described in the present method. This advantage of the GTM becomes clearer for dynamical systems characterized by a high-order differential equations system, say of order  $q$ , where one usually attempts to convert the dynamics into a system of first-order differential equations in some  $q$  unknown variables. Following conventional OC solvers based on spectral methods, and applying full parameterization of the states and the controls, one faces the problem of solving a formidable NLP problem of dimension  $(q(L+1)+M+1)$ , assuming that the control variable is a scalar. Otherwise, the dynamics must be discretized into a system of algebraic equations using higher-order SDMs up to the  $q^{\text{th}}$ -order, where severe ill-conditioning rears its ugly head. Notice here that the GTM still preserves the same dimension  $(L+M+2)$  of the resulting NLP problem. Indeed, many of the complications involved in traditional discretization methods are overcome by the application of the GTM endowed with the P-matrix within the framework of a complete integration numerical scheme, where a significantly small-scale NLP problem and precise solutions can be readily accomplished.

## 7.6 Conclusion and Future Research

This chapter covers a wide collection of CTOCPs with the concrete aim of comparing the efficiency of the current work with other classical discretization methods in the literature. The GTM presented in this chapter is a novel direct optimization method based on Gegenbauer collocation. The method is inspired by our recent achievement in (Elgindy and Smith-Miles, 2013b) in the development of very high-order numerical quadratures based on Gegenbauer polynomials, and represents an extension to the work of Elgindy et al. (2012) to deal with problems where high-order time derivatives of the states arise in the cost function,

## Chapter 7

---

dynamical system, and the path/terminal constraints. To this end, we introduced a substitution  $\mu(t)$  for the highest-order time derivative of the state,  $x^{(N)}(t)$ , and solved the CTOCP directly for  $\mu(t)$  and the control  $u(t)$ . The state vector and its derivatives up to the  $(N - 1)^{\text{th}}$ -order derivative can be accurately evaluated by successive integration. The GTM recasts the dynamical system into its integral form to apply the well-conditioned and stable SIMs, and avoid the ill-conditioning associated with the SDMs. The dynamical system, path and terminal constraints are imposed at the GG points. The integral operations are approximated by optimal Gegenbauer quadratures in the sense of satisfying the optimality constraint (7.7). The successive integrals of the Gegenbauer basis polynomials can be calculated exactly for the GG nodes through the optimal P-matrix. Since the optimal P-matrix is a constant matrix for a particular GG solution points, the GTM can be quickly used to solve many practical trajectory optimization problems, and the Gegenbauer spectral computations can be considerably more effective. It is essential here to acknowledge that the rectangular form of the optimal P-matrix is a useful feature, which allows the GTM to produce higher-order approximations without increasing the dimension of the NLP problem or increasing the number of collocation points. In contrast, traditional spectral methods usually demand that the number of spectral expansion terms  $(N + 1)$  required for the construction of the spectral differentiation/integration matrix be exactly the same as the number of collocation points; cf. (El-Gendi, 1969; Elbarbary, 2007; Fornberg, 1990; Ghoreishi and Hosseini, 2004; Gong et al., 2009; Paraskevopoulos, 1983; Ross and Fahroo, 2002; Weideman and Reddy, 2000). Therefore, to obtain higher-order approximations, one has to increase the size of the spectral matrix, which in turn requires the increase in the number of collocation points. This increase in the number of collocation points is a crucial element in reducing the efficiency of the traditional direct pseudospectral methods and direct orthogonal collocation methods in the sense that:

- (i) the increase of the number of collocation points in a direct pseudospectral method increases the number of unknown spectral coefficients in the state and control expansion series. This fact can be easily derived, since the spectral coefficients in direct pseudospectral methods are exactly the states and controls values at the collocation points, which represent the optimization variables after discretizing the CTOCP into a parameter NLP problem. Hence to achieve higher-order approximations, popular direct pseudospectral methods demand the increase in the dimension of the NLP problem.
- (ii) Although the spectral coefficients of the states and/or controls in direct orthogonal collocation methods may assume any values, and they are not necessarily the same as the states/controls values at the collocation points; cf. (El-Gindy et al., 1995; El-Hawary et al., 2003; Elnagar, 1997; Razzaghi

## Chapter 7

---

and Elnagar, 1993; Vlassenbroeck and Dooren, 1988), the increase in the number of collocation points increases the number of constraints in the resulting NLP problem, since the dynamics and all of the constraints are discretized at the collocation points, and the spectral collocation methods aim for their satisfaction at these particular points.

For a large number ( $M_P + 1$ ) of the Gegenbauer expansion terms, the optimal Gegenbauer quadrature converges to the optimal Chebyshev quadrature in the  $L^\infty$ -norm. Moreover, the optimal Gegenbauer quadrature constructed via Algorithm 2.2 in (Elgindy and Smith-Miles, 2013b) is identical with the Legendre quadrature for large values of  $M_P$  if the approximations are sought in the  $L^2$ -norm. Therefore for collocations of CTOCPs at the Chebyshev-Gauss points, the GTM becomes a direct Chebyshev transcription method, for large numbers of Gegenbauer expansion terms, if the solutions are sought in the  $L^\infty$ -norm. Furthermore, the GTM becomes a direct Legendre transcription method for collocations at the Legendre-Gauss points, for large numbers of the Gegenbauer expansion terms, if the solutions are sought in the  $L^2$ -norm. In fact, due to the precise approximations of the optimal Gegenbauer quadratures adapted in the present GTM, we observed from the extensive numerical experiments that small numbers of collocation points and the states and controls expansion terms are generally sufficient to generate accurate trajectories; cf. Section 7.5. The GTM handles the state and the control constraints smoothly. On the contrary, the presence of such constraints often presents a difficulty in front of the available popular theoretical tools such as Pontryagin's minimum principle and the Hamilton-Jacobi-Bellman optimality equation. Moreover, the GTM approximations converge to the solution of the OC problem much more rapidly than Eulerian approximations. In contrast, Eulerian discretizations require large number of variables, or even experience an explosion in the number of variables to achieve comparable precision of solutions. The significantly small-scale NLP established by the GTM is important to allow real-time decision making as the OC can be readily determined using modern NLP software and computer packages. Moreover, since the GTM solves the CTOCP in the spectral space, the state and control variables can immediately be evaluated at any point in the domain of the solutions. Similar ideas to the ones presented in this chapter can be applied on CTOCPs governed by dynamical systems characterized by integral equations or integro-differential equations; thus the GTM is broadly applicable and encompasses a wider range of problems over the standard direct collocation methods. The robustness and the efficiency of the GTM are verified through four CTOCPs well-studied in the literature. The numerical comparisons conducted in this chapter reveal that the GTM integrated with the optimal Gegenbauer quadrature offers many advantages. Moreover, the results clearly show that the Gegenbauer polynomials are

## Chapter 7

---

effective in direct optimization methods.

One of the useful contributions of this chapter is the establishment of a computationally efficient framework for CTOCPs exhibiting sufficiently differential solutions. Moreover, the current chapter signifies the important advantage of producing very small-scale dimensional NLP problems, which signals the gap between the present method and other traditional methods. The GTM combines the strengths of the versatile Chebyshev, Legendre, and Gegenbauer polynomials in one OC solver, and provides a strong addition to the arsenal of direct optimization methods.

There are many important topics related to the present work, which can be pursued later in the future. In the following, we highlight two important subjects highly relevant to the present work:

- (i) Direct optimization methods based on spectral methods can solve CTOCPs exhibiting nonsmooth/discontinuous solutions with slower convergence rates than those obtained for problems with smooth solutions; cf. (Gong et al., 2006a, Example 5 in pg. 1127) and (Yen and Nagurka, 1991, Example 3 in pg. 212), for instance, for examples on CTOCPs with continuous and discontinuous controllers, respectively. However, the exponential convergence of the present GTM can be easily recovered for CTOCPs with discontinuous/nonsmooth solutions through a “semi-global” approach. The idea is to divide the OC problem into multiple-phases, which can be linked together via continuity conditions (linkage constraints) on the independent variable, the state, and the control. The GTM can then be applied globally within each phase; cf. (Rao, 2003).
- (ii) Further analysis is required to investigate the convergence of the approximate solutions of the GTM based on Gauss collocations to the solutions of the CTOCPs.

### 7.A Elementary Properties and Definitions

In this section we briefly introduce the Gegenbauer polynomials and present some of their useful properties. The Gegenbauer polynomial  $C_n^{(\alpha)}(x)$ ,  $n \in \mathbb{Z}^+$ , of degree  $n$  and associated with the real parameter  $\alpha > -1/2$  is a real-valued function, which appears as an eigensolution to the following singular Sturm-Liouville problem in the finite domain  $[-1, 1]$  (Szegő, 1975):

$$\frac{d}{dx}(1-x^2)^{\alpha+\frac{1}{2}} \frac{dC_n^{(\alpha)}(x)}{dx} + n(n+2\alpha)(1-x^2)^{\alpha-\frac{1}{2}} C_n^{(\alpha)}(x) = 0.$$

## Chapter 7

---

The weight function for the Gegenbauer polynomials is the even function  $(1 - x^2)^{\alpha-1/2}$ . The form of the Gegenbauer polynomials is not unique and depends on a certain standardization. The Gegenbauer polynomials standardized by Doha (1990) so that

$$C_n^{(\alpha)}(x) = \frac{n! \Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} P_n^{(\alpha-\frac{1}{2}, \alpha-\frac{1}{2})}(x), \quad n = 0, 1, 2, \dots, \quad (7.A.1)$$

establish the following useful relations:

$$\begin{aligned} C_n^{(0)}(x) &= T_n(x), \\ C_n^{(\frac{1}{2})}(x) &= L_n(x), \\ C_n^{(1)}(x) &= (1/(n+1)) U_n(x), \end{aligned}$$

where  $P_n^{(\alpha-\frac{1}{2}, \alpha-\frac{1}{2})}(x)$  is the Jacobi polynomial of degree  $n$  and associated with the parameters  $\alpha - \frac{1}{2}, \alpha - \frac{1}{2}$ ;  $T_n(x)$  is the  $n^{\text{th}}$ -degree Chebyshev polynomial of the first kind,  $L_n(x)$  is the  $n^{\text{th}}$ -degree Legendre polynomial;  $U_n(x)$  is the  $n^{\text{th}}$ -degree Chebyshev polynomial of the second type. Throughout the chapter, we shall refer to the Gegenbauer polynomials by those constrained by standardization (7.A.1). The Gegenbauer polynomials are generated by the following Rodrigues' formula:

$$C_n^{(\alpha)}(x) = \left(-\frac{1}{2}\right)^n \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(n + \alpha + \frac{1}{2})} (1 - x^2)^{\frac{1}{2}-\alpha} \frac{d^n}{dx^n} (1 - x^2)^{n+\alpha-\frac{1}{2}}, \quad (7.A.2)$$

or using the three-term recurrence relation

$$(j + 2\alpha)C_{j+1}^{(\alpha)}(x) = 2(j + \alpha)x C_j^{(\alpha)}(x) - j C_{j-1}^{(\alpha)}(x), \quad j \geq 1, \quad (7.A.3)$$

starting from  $C_0^{(\alpha)}(x) = 1$ ;  $C_1^{(\alpha)}(x) = x$ . The set

$$S_n^{(\alpha)} = \{x_j | C_{n+1}^{(\alpha)}(x_j) = 0, j = 0, \dots, n\}, \quad (7.A.4)$$

is the set of the roots/zeros  $\{x_j\}_{j=0}^n$  of the Gegenbauer polynomial  $C_{n+1}^{(\alpha)}(x)$ , which are usually called the GG points. At the special values  $x = \pm 1$ , the Gegenbauer polynomials satisfy the relation

$$C_n^{(\alpha)}(\pm 1) = (\pm 1)^n \quad \forall n. \quad (7.A.5)$$

The interested reader may further pursue more information about the family of Gegenbauer polynomials in many useful textbooks and monographs; cf. (Abramowitz and Stegun, 1965; Andrews et al., 1999; Bayin, 2006; Hesthaven et al., 2007) for instance.

This page is intentionally left blank

## *Chapter 8*

### *Conclusion and Future Research*

# Chapter 8

## Conclusion and Future Research

### 8.1 Summary of Contributions

The numerical methods for solving CTOCPs are conveniently divided into two main categories: IOMs and DOMs. The former class of methods attempts to iterate on the necessary conditions of optimality provided by the CV and the MP to seek their satisfaction while the latter class of methods endeavor to retain the structure of the CTOCPs by discretizing the infinite-dimensional continuous-time problem into a finite dimensional parameter NLP problem, which can be solved by standard optimization software. Although the former class may produce accurate solutions to the CTOCPs, the solution of the resulting TPBVP, as a reformulation of the CTOCP, is usually at least as involved as solving the system equations themselves. The many drawbacks associated with these methods can largely limit their applications on complex OC problems; therefore they are not practical in any but the simplest cases; cf. Chapter 1. In contrast, the theoretical and experimental results of the DOMs presented in this dissertation, in addition to those found in many textbooks, articles, and monographs in the literature favor this class of methods over the former from many perspectives.

The novel DOCM introduced in this dissertation adopts a new vision for the solution of complex CTOCPs based on the simplicity, fast convergence, economy in calculations, and the stability of the Gegenbauer collocation integration schemes. Moreover, one of the major contributions of the presented techniques is the establishment of very small-scale NLP problems analogs to the original CTOCPs, which can deliver very precise solutions for problems with sufficiently differentiable solutions. The foundation for the success of the developed methods largely lies in the precise translation of the integral operations included in the CTOCP without the requirements of either increasing the number of collocation points employed in the discretization of the problem or increasing the dimen-

## Chapter 8

---

sion of the resulting NLP problem. To this end, we have designed an optimal Gegenbauer quadrature which has been thoroughly investigated in Chapter 3. The proposed numerical quadrature is novel in many aspects: (i) The quadrature is established using a rectangular GIM so that the choice of the number of Gegenbauer expansion terms  $(M + 1)$  required for its construction is completely free, and independent of the number of integration nodes. (ii) The quadrature scheme is optimal in the sense that it combines the very useful characteristics of the Chebyshev, Legendre, and Gegenbauer polynomials in one unique quadrature through a unified approach. In particular, the Gegenbauer polynomial expansions are applied in the small/medium range of the spectral expansion terms to produce rapid convergence rates faster than both the Chebyshev and Legendre polynomial expansions. This technique entails the calculation of some optimal Gegenbauer parameter values  $\alpha_i^*$  rather than choosing any arbitrary  $\alpha$  value. For large-scale number of expansion terms, the quadrature is constructed through the elegant Chebyshev and Legendre polynomial expansions, which are optimal in the  $L^\infty$ -norm and  $L^2$ -norm approximations of the smooth functions, respectively. The essential elements discussed in Section 3.2.9 motivated us to develop two efficient computational algorithms, namely Algorithms 2.1 and 2.2, for constructing the novel and optimal P-matrix through interpolations at the optimal set of adjoint GG points (3.15) in the sense of solving Problem (3.13). Algorithm 2.2 is a more cost-effective algorithm suitable for similar sets of integration points, where most of the calculations carried out for the construction of the P-matrix are halved; thus the Gegenbauer spectral computations can be considerably more effective. The construction of the optimal Gegenbauer quadrature is induced by the set of integration points regardless of the integrand function. Moreover, the proposed method establishes a high-order numerical quadrature for any arbitrary sets of integration points, and avoids the Runge Phenomenon through discretizations at the adjoint GG points. The rectangular form of the developed P-matrix is an extremely indispensable element in achieving approximations of higher-orders, and permitting rapid convergence rates without the need to increase the number of integration nodes. The optimal Gegenbauer quadrature is exact for polynomials of any arbitrary degree  $n$  if the number of columns of the P-matrix is greater than or equal to  $n$ . The optimality measure adopted by the present quadrature method, and the applications of the Chebyshev and Legendre polynomial expansions for large-scale number of expansion terms render the optimal Gegenbauer quadrature strong enough to stabilize the calculations, and sufficient to retain the spectral accuracy. The numerical experiments reported in Section 3.3 show that the optimal Gegenbauer quadrature can achieve very rapid convergence rates and higher-order precisions, which can exceed those obtained by the standard Chebyshev and Legendre polynomial methods. Moreover, the optimal Gegenbauer quadrature outperforms conventional Gegenbauer quadrature methods.

## Chapter 8

---

A critical stage in the transcription of a CTOCP by a DOM manifests in the discretization of the dynamics. The caliber of the discrete representation of the dynamical system, and the discrete OC problem in general is crucial for determining the approximate solutions within high accuracy. Chapter 4 highlights the tempting idea of optimizing the GIM to achieve better solution approximations through the transformation of the dynamical system or the OC problem into an unconstrained or constrained optimization problem, respectively. The Gegenbauer parameter  $\alpha$  associated with the Gegenbauer polynomials and employed in the construction of the GIM is then added as an extra unknown variable to be optimized in the resulting optimization problem as an attempt to optimize its value rather than choosing a random value. Although this tempting idea has been applied in a number of articles, Chapter 4 provides a clear and indisputable mathematical proof which rebuffs these approaches in view of the violation of the discrete Gegenbauer orthonormality relation, and the establishment of false optimization problems analogs, which may lead to fallacious solution approximations. The mathematical proof rests upon the fact that the eigenfunctions in spectral theory must be held fixed for defining the projection space, and the approximation procedure must start anew as the space is refined.

Chapter 5 presents an efficient numerical method for discretizing various dynamical systems such as BVPs, integral and integro-differential equations using GIMs. The proposed numerical method avoids the pitfalls of the techniques presented in the preceding chapter, and introduces a strong and practical method for the establishment of approximations of higher-orders to the solutions of continuous dynamical systems. The principle idea is to transform the general BVPs and integro-differential equations into their integral reformulations, which can be discretized efficiently using a hybridization of the GIMs presented in Chapter 3. The resulting algebraic linear system of equations can be solved for the solution values in the physical space using efficient linear system solvers. The proposed hybrid Gegenbauer collocation integration method generally leads to well-conditioned linear systems, and avoid the degradation of precision caused by the severely ill-conditioned SDMs. The theoretical and extensive empirical results presented in this chapter demonstrate the robustness and the spectral accuracy achieved by the proposed method using a relatively small number of solution points. These useful and desirable features are largely due to the rectangular property and the optimality measure adopted by the P-matrix. It has been shown through eight diverse test examples that the performance of the proposed method is superior to other competitive techniques in the recent literature regarding accuracy and convergence rate. Moreover, the developed Gegenbauer collocation integration scheme is memory-minimizing and can be easily programmed.

Chapter 6 presents a novel DOCM using GG collocation for solving CTOCPs with nonlinear dynamics, state and control constraints. The work introduced in

## Chapter 8

---

this chapter represents a major advancement in the area of DOCMs using Gegenbauer polynomials. The proposed GTM transcribes the infinite-dimensional OC problem into a finite-dimensional NLP problem which can be solved in the spectral space; thus approximating the state and the control variables along the entire time horizon. The method was applied on two numerical examples to find the best path in 2D for an unmanned aerial vehicle mobilizing in a stationary risk environment. The implementation of the GTM reveals many fruitful outcomes over the standard variational methods in the sense that: (i) The proposed GTM neither requires the explicit derivation and construction of the necessary conditions nor the calculation of the gradients  $\nabla_x \mathcal{L}$  of the Lagrangian function  $\mathcal{L}(x(t), u(t), t)$  with respect to the state variables, yet it is able to produce rapid convergence and achieve high precision approximations. In contrast, the indirect method applied by Miller et al. (2011) requires the explicit derivation of the adjoint equations, the control equations, and all of the transversality conditions. (ii) To implement a variational method, the user must calculate the gradients  $\nabla_x \mathcal{L}$  for the solution of the necessary conditions of optimality. This property is not a must for a DOM in general. (iii) Since the optimal P-matrix is constant for a particular GG solution points set, the GTM can be quickly used to solve many practical trajectory optimization problems. We observed also that decreasing the number of columns  $M$  of the P-matrix and the parameter  $M_{\max}$  required for the construction of the P-matrix via Algorithm 2.2 presented in Chapter 3 can reduce the calculations time taken by the GTM for solving CTOCPs with a slight reduction in accuracy. For instance, in Example 6.4.1, the drop in the values of  $M = M_{\max}$  by 6 units, and under similar parameter settings, results in a slight change of 0.0418 in the computed risk integral value  $J$ . The recorded average CPU time in 100 runs taken by the GTM in this case was shorter by 0.5924 seconds. On the other hand, to carry out a variational method, one usually bears the labor of constructing the necessary conditions of optimality offline before the start of the optimization process on the digital computer. Even if the user managed to determine the necessary conditions of optimality online using symbolic arithmetic, the latter can be too slow in practice (Keyser et al., 1998; Krishnan and Manocha, 1995). (iv) The GTM is in principle robust, and tends to have better convergence than the variational methods, which lead to unstable TPBVPs with very small radii of convergence. Another notable advantage of the GTM is that the successive integrals of the Gegenbauer basis polynomials can be calculated exactly at the GG points through the optimal P-matrix; thus the numerical error arises due to the round-off errors and the fact that a finite number of the Gegenbauer basis polynomials are used to represent the state and the control variables. (v) The GTM handles the system dynamics using spectral integration matrices (SIMs) celebrated for their stability and well-conditioning rather than the SDMs which suffer from severe ill-conditioning, and are prone to large round-off errors. In con-

## Chapter 8

---

trast, typical variational techniques carried out using spectral methods discretize the linear TPBVP into a system of algebraic equations by way of a differentiation matrix; cf. (Yan et al., 2001) for instance. (vi) The GTM deals with the state and the control constraints smoothly; on the contrary, the presence of such constraints often presents a difficulty in front of the popular classical theoretical tools such as the MP and the HJB equation. (vii) The developed GTM can be easily extended to higher dimensional OC problems under the same level of complexity, whereas the difficulty of establishing the necessary conditions stands as a coarse barrier against the extension of the variational methods to more complex OC problems.

Chapter 7 covers a wider collection of CTOCPs with the concrete aim of comparing the efficiency of the GTM with other classical discretization methods in the literature. The GTM presented in this chapter extends the work introduced in the previous chapter to deal with problems where different orders time derivatives of the states arise in the cost function, dynamical system, and the path/terminal constraints. To this end, we introduced a substitution  $\mu(t)$  for the highest-order time derivative of the state,  $x^{(N)}(t)$ ,  $N \in \mathbb{Z}^+$ , and solved the CTOCP directly for  $\mu(t)$  and the control  $u(t)$ . The state vector and its derivatives up to the  $(N - 1)$ th-order derivative can be accurately evaluated by successive integration. This key idea provides the luxury of working in a full integration environment, taking full advantage of the well-stability of the integral operators. Moreover, we investigated the solution of LQR problems characterized by linear time-invariant dynamical systems. The GTM outperforms DLCMs in many aspects, but mainly: (i) The exponential convergence of the GTM clearly observed through Tables 7.1–7.4 shows an advantage over classical direct local collocation schemes based on Eulerian like discretizations, which require large number of variables, or even experience an explosion in the number of variables to achieve comparable precision of solutions. (ii) Since the GTM solves the CTOCP in the spectral space, the state and control variables can immediately be evaluated at any point in the domain of the solutions. In contrast, typical DLCMs based on finite-difference schemes, for instance, invoke an interpolation method to evaluate the state and control histories at an intermediate solution point. On the other hand, the extensive numerical comparisons conducted in this chapter signifies the superiority of the proposed GTM over traditional DOCMs and direct PS methods regarding robustness, accuracy, economy in calculations, rates of convergence, and the production of significantly small-scale NLP problems. The wide gap between the proposed GTM and other discretization methods is largely due to the following key elements: (i) The rectangular form of the optimal P-matrix allows the GTM to produce higher-order approximations without increasing the dimension of the NLP problem or increasing the number of collocation points. In contrast, the popular direct PS methods demand the increase

## Chapter 8

---

in the size of the SDM to achieve higher-order approximations, which in turn requires the increase in the number of collocation points and the dimension of the NLP problem. In particular, one usually cannot obtain higher-order approximations in a PS method without increasing the size of each of these three key elements, namely, (a) the size of the SDM, (b) the number of collocation points, and (c) the number of state and control expansion terms. (ii) The GTM works under a complete integration environment through the recast of the dynamics into its integral formulation. The integral form dynamics is then discretized using the well-conditioned SIMs. In this manner, the proposed GTM unifies the process of the discretization of the dynamics and the integral cost function. On the contrary, classical DOCMs and direct PS methods discretize the dynamical system using standard square SDMs, which are associated with many drawbacks. (iii) The GTM employs Gegenbauer polynomial expansions for the small/medium range of the number of spectral expansion terms to produce higher-order approximations. For a large number of the Gegenbauer expansion terms ( $M + 1$ ), the optimal Gegenbauer quadrature converges to the optimal Chebyshev quadrature in the  $L^\infty$ -norm. Moreover, the optimal Gegenbauer quadrature constructed via Algorithm 2.2 in Chapter 3 is identical with the Legendre quadrature for large values of  $M > M_{\max}$  if the approximations are sought in the  $L^2$ -norm. Therefore for collocations of CTOCPs at the Chebyshev-Gauss points, the GTM becomes a direct Chebyshev transcription method, for large numbers of the Gegenbauer expansion terms, if the solutions are sought in the  $L^\infty$ -norm. Furthermore, the GTM becomes a direct Legendre transcription method for collocations at the LG points, for large numbers of the Gegenbauer expansion terms, if the solutions are sought in the  $L^2$ -norm. In fact, due to the precise approximations of the optimal Gegenbauer quadratures adapted in the present GTM, we observed from the extensive numerical experiments conducted in Chapters 6 and 7 that small numbers of the collocation points and the states and controls expansion terms are generally sufficient to generate very accurate trajectories. Hence the GTM combines the strengths of the versatile Chebyshev, Legendre, and Gegenbauer polynomials in one OC solver, and provides a strong addition to the arsenal of DOMs. On the contrary, typical DOCMs and direct PS methods apply Chebyshev polynomial expansions or Legendre polynomial expansions blindly for all types of expansions and any type of OC problems.

The significantly smaller-scale NLP problem established by the GTM introduced in this dissertation is of importance to allow real-time decision making as the OC can be readily determined using modern NLP software and computer packages; cf. Table 7.5 for instance. The robustness and the efficiency of the GTM are verified through extensive well-studied CTOCPs in the literature. The numerical comparisons conducted in this dissertation reveal that the Gegenbauer collocation integration schemes integrated with the optimal Gegenbauer quadra-

## Chapter 8

---

ture offer many advantages, and yield a better control performance compared to other conventional computational OC methods. Moreover, the results clearly show that the Gegenbauer polynomials are very effective in DOMs.

One of the paramount contributions of this dissertation is the establishment of computationally very efficient frameworks for solving dynamical systems and CTOCPs exhibiting sufficiently differential solutions. The introduced ideas present major breakthroughs in the areas of dynamical systems and computational OC theory as they deliver significantly accurate solutions using considerably small numbers of collocation points. Moreover, the dissertation signifies the very important advantage of producing very small-scale dimensional NLP problems, and highlights the gap between the present GTM and other traditional methods. Since the Gegenbauer collocation method can provide excellent approximations to the integration in the cost function (as extensively studied in Chapter 3), the differential/integral equations of the dynamical system (as comprehensively investigated in Chapter 5), and the state-control constraints (as evident from Chapters 6 and 7), it is the method of choice for many types of mathematical problems, and well suited for digital computations.

### 8.2 Future Research Directions

Numerical strategies and optimization techniques yielding fast and accurate approximations to the solutions are highly desirable to allow real-time decision making. In fact, there is a number of intriguing research points which may be pursued later. In the following, we mention some of them:

- (i) A nonlinear OC problem may admit multiple solutions; cf. (Ghosh et al., 2011; Kogut and Leugering, 2011; Singh, 2010). Therefore, one may apply global optimization solvers such as the genetic algorithms, evolution strategies, particle swarm optimization, ant colony optimization, etc. for the minimization of the resulting NLP problems instead of the “fmincon” MATLAB local optimization solver applied in this dissertation.
- (ii) Similar ideas to the methods described in this dissertation can be easily extended to solve CTOCPs governed by integral equations or integro-differential equations; therefore, the GTM encompasses a wider range of OC problems over the standard DOMs.
- (iii) The mathematical convergence proof of the Gegenbauer collocation integration method for solving TPBVPs is provided in Chapter 5. Moreover, we notice from our empirical experience in solving several problems with

## Chapter 8

---

analytic solutions in this dissertation using the proposed algorithms that the presented Gegenbauer collocation integration methods converge very rapidly to the sought solutions as the number of collocation points, state and control expansion terms increase. However, at this stage, we do not have mathematical convergence proofs for the proposed algorithms for the solution of general dynamical systems and CTOCPs. Therefore, further tests and analysis are necessary to investigate the stability, the accuracy, and the convergence of the Gegenbauer collocation integration methods presented in this dissertation.

- (iv) The GTM is significantly more accurate than other conventional direct local methods for smooth OC problems, enjoying the so called “spectral accuracy.” Moreover, Chapter 7 highlights the significant advantages of the GTM over the DOCMs and direct PS methods. For the class of discontinuous/nonsmooth OC problems, the existence and convergence results of the similar approaches of direct PS methods have been investigated and proved in a number of articles; cf. (Kang et al., 2005, 2007, 2008), for instances, for studies on OC problems with discontinuous controller using Legendre polynomials. Here it is essential to acknowledge that the convergence rate of standard DOCMs/PS methods applied for discontinuous/nonsmooth OC problems is not imposing as clearly observed for OC problems with smooth solutions. In fact, the superior accuracy of the GTM cannot be realized in the presence of discontinuities and/or nonsmoothness in the OC problem, or in its solutions, as the convergence rate grows slower in this case for increasing number of GG collocation points and Gegenbauer expansion terms. Some research studies in this area manifest that the accuracies of DGCMS and DLCMS become comparable for nonsmooth OC problems; cf. (Huntington, 2007). To recover the exponential convergence property of the GTM in the latter case, the GTM can be applied within the framework of a semi-global approach. Here the OC problems can be divided into multiple-phases, which can be linked together via continuity conditions (linkage constraints) on the independent variable, the state, and the control. The GTM can then be applied globally within each phase. The reader may consult Ref. (Rao, 2003), for instance, for a similar practical implementation of this solution method. Another possible approach to accelerate the convergence rate of the GTM, and to recover the spectral accuracy, is to treat the GTM with an appropriate smoothing filter; cf. (Elnagar and Kazemi, 1998b), for instance, for a parallel approach using a PS Legendre method. Other methods include the knotting techniques developed in (Ross and Fahroo, 2002, 2004) for solving nonsmooth OC problems, where the dynamics are governed by controlled differential inclusions. Moreover, the Gegenbauer

## Chapter 8

---

reconstruction method invented by Gottlieb et al. in 1992 is another potential method which may be extended for solving OC problems exhibiting nonsmooth/discontinuous solutions.

- (v) The numerical methods developed in this dissertation assume global smoothness, and generally use a single grid for discretization. An interesting direction for future work could involve a study of composite Gegenbauer grids and adaptivity to improve the convergence behavior of the solvers for complex problems.

## References

- Abramowitz, M., Stegun, I. A., 1965. Handbook of Mathematical Functions. Dover.
- Ahmadi, A., Green, T. C., July 2009. Optimal power flow for autonomous regional active network management system. In: Power Energy Society General Meeting, 2009. PES '09. IEEE. pp. 1–7.
- Ahmed, S., Muldoon, M., Spigler, R., 1986. Inequalities and numerical bounds for zeros of ultraspherical polynomials. *SIAM Journal on Mathematical Analysis* 17 (4), 1000–1007.
- Ait, W., Mackenroth, U., 1989. Convergence of finite element approximation to state constrained convex parabolic boundary control problems. *SIAM Journal on Control and Optimization* 27 (4), 718–736.
- Anderson, B. D. O., Moore, J. B., 2007. Optimal Control: Linear Quadratic Methods. Dover Books on Engineering Series. Dover Publications.
- Andreev, M. A., Miller, A. B., Miller, B. M., Stepanyan, K. V., 2012. Path planning for unmanned aerial vehicle under complicated conditions and hazards. *Journal of Computer and Systems Sciences International* 51, 328–338.
- Andrews, G. E., Askey, R., Roy, R., 1999. Special Functions. Cambridge University Press, Cambridge.
- Aoki, M., September 1960. Dynamic programming approach to a final-value control system with a random variable having an unknown distribution function. *IRE Transactions on Automatic Control* 5 (4), 270–283.
- Apreutesei, N. C., 2012. An optimal control problem for a pest, predator, and plant system. *Nonlinear Analysis: Real World Applications* 13 (3), 1391–1400.
- Archibald, R., Chen, K., Gelb, A., Renaut, R., September 2003. Improving tissue segmentation of human brain MRI through preprocessing by the Gegenbauer reconstruction method. *NeuroImage* 20 (1), 489–502.
- Area, I., Dimitrov, D. K., Godoy, E., Ronveaux, A., 23 March 2004. Zeros of Gegenbauer and Hermite polynomials and connection coefficients. *Mathematics of Computation* 73 (248), 1937–1951.
- Atkinson, K. E., 1997. The Numerical Solution of Integral Equation of the Second Kind. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge.

- Babolian, E., Fattahzadeh, F., 2007. Numerical solution of differential equations by using Chebyshev wavelet operational matrix of integration. *Applied Mathematics and Computation* 188 (1), 417–426.
- Balachandran, B., Kalmár-Nagy, T., Gilsinn, D. E., 2009. *Delay Differential Equations: Recent Advances and New Directions*. Springer.
- Barranco, J. A., Marcus, P. S., 2006. Multidomain local Fourier method for PDEs in complex geometries. *Journal of Computational Physics* 219 (1), 21–46.
- Barrio, R., 1999. On the A-stability of Runge-Kutta collocation methods based on orthogonal polynomials. *SIAM Journal on Numerical Analysis* 36 (4), 1291–1303.
- Bassey, K. J., Chigbu, P. E., 2012. On optimal control theory in marine oil spill management: A markovian decision approach. *European Journal of Operational Research* 217 (2), 470–478.
- Bavinck, H., Hooghiemstra, G., Waard, E. D., 1993. An application of Gegenbauer polynomials in queueing theory. *Journal of Computational and Applied Mathematics* 49 (1–3), 1–10.
- Bayin, Ş. S., 18 July 2006. *Mathematical Methods in Science and Engineering*. Wiley-Interscience.
- Becerra, V. M., 2011. PSOPT Optimal Control Solver User Manual - Release 3 build 2011-07-28. Available: <http://www.psopt.org/>.
- Becker, R., Vexler, B., 2007. Optimal control of the convection-diffusion equation using stabilized finite element methods. *Numerische Mathematik* 106 (3), 349–367.
- Bellman, R., 2003. *Dynamic Programming*. Dover Books on Mathematics. Dover Publications.
- Ben-yu, G., 1998. Gegenbauer approximation and its applications to differential equations on the whole line. *Journal of Mathematical Analysis and Applications* 226 (1), 180–206.
- Benson, D., 2005. *A Gauss Pseudospectral Transcription for Optimal Control*. Ph.D. thesis, Massachusetts Institute of Technology, United States–Massachusetts.
- Benson, D. A., 2004. *A Gauss Pseudospectral Transcription for Optimal Control*. Ph.D. dissertation, MIT.

- Benson, D. A., Huntington, G. T., Thorvaldsen, T. P., Rao, A. V., 2006. Direct trajectory optimization and costate estimation via an orthogonal collocation method. *Journal of Guidance, Control, and Dynamics* 29 (6), 1435–1440.
- Bernardi, C., Maday, Y., 1997. Spectral methods, *Handbook of Numerical Analysis*. Vol. 5 of *Techniques of Scientific Computing (Part 2)*. P.G. Ciarlet; J.L. Lions eds., North-Holland.
- Bertsekas, D. P., 2005. *Dynamic Programming and Optimal Control*, 3rd Edition. Vol. I of *Athena Scientific optimization and computation series*. Athena Scientific.
- Bertsekas, D. P., 2007. *Dynamic Programming and Optimal Control*. Vol. 2 of *Athena Scientific Optimization and Computation Series*. Athena Scientific.
- Betts, J. T., 1998. Survey of numerical methods for trajectory optimization. *Journal of Guidance, Control, and Dynamics* 21 (2), 193–207.
- Betts, J. T., 2001. *Practical Methods for Optimal Control Using Nonlinear Programming*. SIAM, Philadelphia.
- Betts, J. T., December 2009. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*, 2nd Edition. No. 19 in *Advances in Design and Control*. SIAM, Philadelphia.
- Bock, H. G., Plitt, K. J., 1984. A multiple shooting algorithm for direct solution of optimal control problems. In: *Proceedings of the 9th IFAC World Congress*. Budapest, pp. 242–247.
- Boltyanskii, V. G., Gamkrelidze, R. V., Pontryagin, L. S., 1956. Towards a theory of optimal processes (Russian). *Reports Acad. Sci. USSR* 110 (1).
- Bonnans, J. F., Laurent-Varin, J., 2006. Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control. *Numerische Mathematik* 103, 1–10.
- Borzi, A., Von Winckel, G., 2009. Multigrid methods and sparse-grid collocation techniques for parabolic optimal control problems with random coefficients. *SIAM Journal on Scientific Computing* 31 (3), 2172–2192.
- Boyd, J. P., 1989. *Chebyshev and Fourier Spectral Methods*. Springer-Verlag, Berlin.
- Boyd, J. P., 2000. *Chebyshev and Fourier Spectral Methods*, 2nd Edition. Dover, NY.

- Boyd, J. P., 2001. Chebyshev and Fourier Spectral Methods. Dover Books on Mathematics Series. Dover Publications.
- Boyd, J. P., 2006. Computing the zeros, maxima and inflection points of Chebyshev, Legendre and Fourier series: solving transcendental equations by spectral interpolation and polynomial rootfinding. *Journal of Engineering Mathematics* 56 (3), 203–219.
- Breuer, K. S., Everson, R. M., 1992. On the errors incurred calculating derivatives using Chebyshev polynomials. *Journal of Computational Physics* 99 (1), 56–67.
- Bryson, A. E., Ho, Y. C., 1975. *Applied Optimal Control: Optimization, Estimation, and Control*. Halsted Press Book. Taylor & Francis.
- Bryson, A.E., J., June 1996. Optimal control-1950 to 1985. *IEEE Control Systems* 16 (3), 26–33.
- Burden, R. L., Faires, J. D., 2000. *Numerical Analysis*, 7th Edition. Brooks Cole.
- Butkovsky, A., Egorov, A., Lurie, K., 1968. Optimal control of distributed systems (a survey of Soviet publications). *SIAM Journal on Control* 6 (3).
- Canuto, C., Hussaini, M. Y., Quarteroni, A., Zang, T. A., 1988. *Spectral Methods in Fluid Dynamics*. Springer-Verlag, Berlin.
- Canuto, C., Hussaini, M. Y., Quarteroni, A., Zang, T. A., 2006. *Spectral Methods: Fundamentals in Single Domains*. Scientific Computation. Springer, Berlin; NY.
- Canuto, C., Hussaini, M. Y., Quarteroni, A., Zang, T. A., 2007. *Spectral Methods: Evolution to Complex Geometries and Applications to Fluid Dynamics*. Springer-Verlag, Berlin; New York.
- Cattani, C., 2008. Shannon wavelets theory. *Mathematical Problems in Engineering* 2008, 1–24.
- Chakraborty, K., Das, S., Kar, T. K., 2011. Optimal control of effort of a stage structured prey-predator fishery model with harvesting. *Nonlinear Analysis* 12 (6), 3452–3467.
- Chen, C., Hsiao, C., October 1975. Design of piecewise constant gains for optimal control via Walsh functions. *IEEE Transactions on Automatic Control* 20 (5), 596–603.

- Chen, Y., Huang, F., Yi, N., Liu, W., 2011. A Legendre-Galerkin spectral method for optimal control problems governed by Stokes equations. *SIAM Journal on Numerical Analysis* 49 (4), 1625–1648.
- Chen, Y., Lu, Z., November 2010. Error estimates for parabolic optimal control problem by fully discrete mixed finite element methods. *Finite Elements in Analysis and Design* 46 (11), 957–965.
- Cheney, E. W., Kincaid, D. R., 2012. *Numerical Mathematics and Computing*. Cengage Learning.
- Chou, J., Horng, I., 1985a. Application of Chebyshev polynomials to the optimal control of time-varying linear systems. *International Journal of Control* 41 (1), 135–144.
- Chou, J., Horng, I., 1985b. Shifted Chebyshev series analysis of linear optimal control systems incorporating observers. *International Journal of Control* 41 (1), 129–134.
- Clenshaw, C. W., 1957. The numerical solution of linear differential equations in Chebyshev series. *Mathematical Proceedings of the Cambridge Philosophical Society* 53 (1), 134–149.
- Clenshaw, C. W., Curtis, A. R., 1960. A method for numerical integration on an automatic computer. *Numerische Mathematik* 2, 197–205.
- Coutsias, E. A., Hagstrom, T., Hesthaven, J. S., Torres, D., 1996a. Integration preconditioners for differential operators in spectral  $\tau$ -methods. In: *Proceedings of the Third International Conference on Spectral and High Order Methods*. A.V. Ilin, L.R. Scott (Eds.), Houston, TX, pp. 21–38.
- Coutsias, E. A., Hagstrom, T., Torres, D., 1996b. An efficient spectral method for ordinary differential equations with rational function coefficients. *Mathematics of Computation* 65 (214), 611–635.
- Coverstone-Carroll, V. L., Wilkey, N. M., 1995. Optimal control of a satellite-robot system using direct collocation with non-linear programming. *Acta Astronautica* 36 (3), 149–162.
- Cushman-Roisin, B., Beckers, J., 2011. *Introduction to Geophysical Fluid Dynamics: Physical and Numerical Aspects*, 2nd Edition. Vol. 101 of *International Geophysics Series*. Academic Press.

- Cuthrell, J. E., Biegler, L. T., 1989. Simultaneous optimization and solution methods for batch reactor control profiles. *Computers & Chemical Engineering* 13 (1-2), 49-62.
- Danfu, H., Xufeng, S., 2007. Numerical solution of integro-differential equations by using CAS wavelet operational matrix of integration. *Applied Mathematics and Computation* 194 (2), 460-466.
- Darby, C. L., 2011. hp-Pseudospectral Method for Solving Continuous-Time Non-linear Optimal Control Problems. Ph.D. dissertation, University of Florida.
- Darby, C. L., Hager, W. W., Rao, A. V., 2011. An hp-adaptive pseudospectral method for solving optimal control problems. *Optimal Control Applications and Methods* 32 (4), 476-502.
- Deissenberg, C., Hartl, R. F., 2005. *Optimal Control and Dynamic Games: Applications in Finance, Management Science and Economics*. Advances in computational management science. Springer.
- Delves, L. M., Mohamed, J. L., 1985. *Computational Methods for Integral Equations*. Cambridge.
- Deshpande, S. A., Agashe, S. D., 2011. Application of a parametrization method to problem of optimal control. *Journal of the Franklin Institute* 348 (9), 2390-2405.
- Ding, Y., Wang, S. S. Y., 2012. Optimal control of flood water with sediment transport in alluvial channel. *Separation and Purification Technology* 84, 85-94.
- Doha, E., Abd-Elhameed, W., 2002. Efficient spectral-Galerkin algorithms for direct solution of second-order equations using ultraspherical polynomials. *SIAM Journal on Scientific Computing* 24 (2), 548-571.
- Doha, E. H., 1990. An accurate solution of parabolic equations by expansion in ultraspherical polynomials. *Computers & Mathematics with Applications* 19 (4), 75-88.
- Doha, E. H., 1991. The coefficients of differentiated expansions and derivatives of ultraspherical polynomials. *Computers & Mathematics with Applications* 21 (2-3), 115-122.

- Doha, E. H., 15 February 2002. On the coefficients of integrated expansions and integrals of ultraspherical polynomials and their applications for solving differential equations. *Journal of Computational and Applied Mathematics* 139 (2), 275–298.
- Doha, E. H., Abd-Elhameed, W. M., July 2009. Efficient spectral ultraspherical-dual-Petrov-Galerkin algorithms for the direct solution of  $(2n + 1)$ th-order linear differential equations. *Mathematics and Computers in Simulation* 79 (11), 3221–3242.
- Don, W. S., Solomonoff, A., 1997. Accuracy enhancement for higher derivatives using Chebyshev collocation and a mapping technique. *SIAM Journal on Scientific Computing* 18 (4), 1040–1055.
- Dontchev, A. L., Hager, W. W., 1997. The Euler approximation in state constrained optimal control. *Mathematics of Computation* 70 (233), 173–203.
- Dontchev, A. L., Hager, W. W., Veliov, V. M., 2000. Second-order Runge–Kutta approximations in control constrained optimal control. *SIAM Journal on Numerical Analysis* 38 (1), 202–226.
- Driscoll, T. A., 20 August 2010. Automatic spectral collocation for integral, integro-differential, and integrally reformulated differential equations. *Journal of Computational Physics* 229 (17), 5980–5998.
- Driscoll, T. A., Bornemann, F., Trefethen, L. N., 2008. The chebop system for automatic solution of differential equations. *BIT Numerical Mathematics* 48, 701–723.
- Dzhuraev, A., 1992. *Methods of Singular Integral Equations*. Longman Scientific & Technical, London; NY.
- El-Gendi, S. E., 1969. Chebyshev solution of differential, integral, and integro-differential equations. *Computer Journal* 12 (3), 282–287.
- El-Gindy, T. M., El-Hawary, H. M., Salim, M. S., El-Kady, M., March 1995. A Chebyshev approximation for solving optimal control problems. *Computers & Mathematics with Applications* 29 (6), 35–45.
- El-Gindy, T. M., Salim, M. S., 1990. Penalty functions with partial quadratic interpolation technique in the constrained optimization problems. *Journal of Institute of Mathematics & Computer Sciences* 3 (1), 85–90.

- El-Hawary, H. M., Salim, M. S., Hussien, H. S., 2000. An optimal ultraspherical approximation of integrals. *International Journal of Computer Mathematics* 76 (2), 219–237.
- El-Hawary, H. M., Salim, M. S., Hussien, H. S., 2003. Ultraspherical integral method for optimal control problems governed by ordinary differential equations. *Journal of Global Optimization* 25 (3), 283–303.
- El-Kady, M. M., Hussien, H. S., Ibrahim, M. A., 2009. Ultraspherical spectral integration method for solving linear integro-differential equations. *World Academy of Science, Engineering and Technology* 33, 880–887.
- Elbarbary, E. M. E., 2006. Integration preconditioning matrix for ultraspherical pseudospectral operators. *SIAM Journal on Scientific Computing* 28 (3), 1186–1201.
- Elbarbary, E. M. E., 2007. Pseudospectral integration matrix and boundary value problems. *International Journal of Computer Mathematics* 84 (12), 1851–1861.
- Elgindy, K. T., 2008. Chebyshev Approximation for Solving Differential Equations, Integral Equations and Nonlinear Programming Problems. M.Sc. thesis, Assiut University, Assiut, Egypt.
- Elgindy, K. T., 15 March 2009. Generation of higher order pseudospectral integration matrices. *Applied Mathematics and Computation* 209 (2), 153–161.
- Elgindy, K. T., Hedar, A., 15 December 2008. A new robust line search technique based on Chebyshev polynomials. *Applied Mathematics and Computation* 206 (2), 853–866.
- Elgindy, K. T., Smith-Miles, K. A., 15 October 2013a. Fast, accurate, and small-scale direct trajectory optimization using a Gegenbauer transcription method. *Journal of Computational and Applied Mathematics* 251 (0), 93–116.
- Elgindy, K. T., Smith-Miles, K. A., April 2013b. Optimal Gegenbauer quadrature over arbitrary integration nodes. *Journal of Computational and Applied Mathematics* 242 (0), 82–106.
- Elgindy, K. T., Smith-Miles, K. A., 1 January 2013c. Solving boundary value problems, integral, and integro-differential equations using Gegenbauer integration matrices. *Journal of Computational and Applied Mathematics* 237 (1), 307–325.

- Elgindy, K. T., Smith-Miles, K. A., Miller, B., 15–16 November 2012. Solving optimal control problems using a Gegenbauer transcription method. In: The Proceedings of 2012 Australian Control Conference, AUCC 2012. Engineers Australia, University of New South Wales, Sydney, Australia.
- Elnagar, G., Kazemi, M., Razzaghi, M., 1995. The pseudospectral Legendre method for discretizing optimal control problems. *IEEE Transactions on Automatic Control* 40 (10), 1793–1796.
- Elnagar, G., Zafiris, V., 2005. A Chebyshev spectral method for time-varying two-point boundary-value and optimal control problems. *International Journal of Computer Mathematics* 82 (2), 193–202.
- Elnagar, G. N., 3 March 1997. State-control spectral Chebyshev parameterization for linearly constrained quadratic optimal control problems. *Journal of Computational and Applied Mathematics* 79 (1), 19–40.
- Elnagar, G. N., Kazemi, M. A., 1995. A cell-averaging Chebyshev spectral method for the controlled Duffing oscillator. *Applied Numerical Mathematics* 18 (4), 461–471.
- Elnagar, G. N., Kazemi, M. A., 1998a. Pseudospectral Chebyshev optimal control of constrained nonlinear dynamical systems. *Computational Optimization and Applications* 11, 195–217.
- Elnagar, G. N., Kazemi, M. A., 1998b. Pseudospectral Legendre-based optimal computation of nonlinear constrained variational problems. *Journal of Computational and Applied Mathematics* 88 (2), 363–375.
- Elnagar, G. N., Razzaghi, M., 1997. A collocation-type method for linear quadratic optimal control problems. *Optimal Control Applications and Methods* 18 (3), 227–235.
- Endow, Y., July 1989. Optimal control via Fourier series of operational matrix of integration. *IEEE Transactions on Automatic Control* 34 (7), 770–773.
- Engelhart, M., Lebedz, D., Sager, S., 2011. Optimal control for selected cancer chemotherapy ODE models: A view on the potential of optimal schedules and choice of objective function. *Mathematical Biosciences* 229 (1), 123–134.
- Enright, P. J., Conway, B. A., July–August 1992. Discrete approximations to optimal trajectories using direct transcription and nonlinear programming. *Journal of Guidance, Control, and Dynamics* 15 (4), 994–1002.

- Fahroo, F., Ross, I. M., 2000. Direct trajectory optimization by a Chebyshev pseudospectral method. In: Proceedings of the 2000 American Control Conference. Vol. 6. pp. 3860–3864.
- Fahroo, F., Ross, I. M., 2001. Costate estimation by a Legendre pseudospectral method. *Journal of Guidance, Control, and Dynamics* 24 (2), 270–277.
- Fahroo, F., Ross, I. M., 2002. Direct trajectory optimization by a Chebyshev pseudospectral method. *Journal of Guidance, Control, and Dynamics* 25 (1), 160–166.
- Fahroo, F., Ross, I. M., 2008. Pseudospectral methods for infinite-horizon optimal control problems. *Journal of Guidance, Control, and Dynamics* 31 (4), 927–936.
- Fornberg, B., 1990. An improved pseudospectral method for initial-boundary value problems. *Journal of Computational Physics* 91 (2), 381–397.
- Fornberg, B., 1996. *A Practical Guide to Pseudospectral Methods*. Vol. 1 of Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge.
- Foroozandeh, Z., Shamsi, M., March–April 2012. Solution of nonlinear optimal control problems by the interpolating scaling functions. *Acta Astronautica* 72, 21–26.
- Fraser-Andrews, G., 1999. A multiple-shooting technique for optimal control. *Journal of Optimization Theory and Applications* 102, 299–313.
- Funaro, D., 1987. A preconditioning matrix for the Chebyshev differencing operator. *SIAM Journal on Numerical Analysis* 24 (5), 1024–1031.
- Funaro, D., 1992. *Polynomial Approximations of Differential Equations*. Springer-Verlag.
- Gao, Y., Li, X., January 2010. Optimization of low-thrust many-revolution transfers and Lyapunov-based guidance. *Acta Astronautica* 66 (1–2), 117–129.
- Gardner, D. R., Trogdon, S. A., Douglass, R. W., 1989. A modified tau spectral method that eliminates spurious eigenvalues. *Journal of Computational Physics* 80 (1), 137–167.
- Garg, D., August 2011. *Advances in Global Pseudospectral Methods for Optimal Control*. Ph.D. thesis, University of Florida.

- Garg, D., Hager, W. W., Rao, A. V., 2011a. Pseudospectral methods for solving infinite-horizon optimal control problems. *Automatica* 47 (4), 829–837.
- Garg, D., Patterson, M. A., Francolin, C., Darby, C. L., Huntington, G. T., Hager, W. W., Rao, A. V., 2011b. Direct trajectory optimization and costate estimation of finite-horizon and infinite-horizon optimal control problems using a Radau pseudospectral method. *Computational Optimization and Applications* 49 (2), 335–358.
- Garg, D., Patterson, M. A., Hager, W. W., Rao, A. V., Benson, D. A., Huntington, G. T., 2010. A unified framework for the numerical solution of optimal control problems using pseudospectral methods. *Automatica* 46 (11), 1843–1851.
- Gelb, A., 2004. Parameter optimization and reduction of round off error for the Gegenbauer reconstruction method. *Journal of Scientific Computing* 20 (3), 433–459.
- Gelb, A., Gottlieb, S., 2007. The resolution of the Gibbs phenomenon for Fourier spectral methods. In: Jerri, A. (Ed.), *Advances in The Gibbs Phenomenon*. Sampling Publishing, Potsdam, New York.
- Gelb, A., Jackiewicz, Z., 2005. Determining analyticity for parameter optimization of the Gegenbauer reconstruction method. *SIAM Journal on Scientific Computing* 27 (3), 1014–1031.
- Ghoreishi, F., Hosseini, S. M., 2004. A preconditioned implementation of pseudospectral methods on arbitrary grids. *Applied Mathematics and Computation* 148 (1), 15–34.
- Ghoreishi, F., Hosseini, S. M., 15 April 2008. Integration matrix based on arbitrary grids with a preconditioner for pseudospectral method. *Journal of Computational and Applied Mathematics* 214 (1), 274–287.
- Ghosh, A., Das, S., Chowdhury, A., Giri, R., 2011. An ecologically inspired direct search method for solving optimal control problems with Bézier parameterization. *Engineering Applications of Artificial Intelligence* 24 (7), 1195–1203.
- Gill, P., Murray, W., Saunders, M., 2005. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM Review* 47 (1), 99–131.
- Glabisz, W., January 2004. The use of Walsh-wavelet packets in linear boundary value problems. *Computers & Structures* 82 (2–3), 131–141.

- Goh, C. J., Teo, K. L., 1988. Control parametrization: A unified approach to optimal control problems with general constraints. *Automatica* 24 (1), 3–18.
- Golberg, M. A., 1990. *Numerical Solution of Integral Equations*. Plenum, NY.
- Gong, Q., Kang, W., Ross, I., July 2006a. A pseudospectral method for the optimal control of constrained feedback linearizable systems. *IEEE Transactions on Automatic Control* 51 (7), 1115–1129.
- Gong, Q., Ross, I., Kang, W., Fahroo, F., December 2006b. On the pseudospectral covector mapping theorem for nonlinear optimal control. In: *45th IEEE Conference on Decision and Control*. pp. 2679–2686.
- Gong, Q., Ross, I. M., Fahroo, F., December 2009. A Chebyshev pseudospectral method for nonlinear constrained optimal control problems. In: *Proceedings of the 48th IEEE Conference on Decision and Control, held jointly with the 28th Chinese Control Conference. CDC/CCC 2009*. pp. 5057–5062.
- Gong, Q., Ross, I. M., Kang, W., 11–13 July 2007. A unified pseudospectral framework for nonlinear controller and observer design. In: *American Control Conference, 2007. ACC '07. Marriott Marquis Hotel at Times Square, NY, USA*, pp. 1943–1949.
- Gong, Q., Ross, I. M., Kang, W., Fahroo, F., 2008. Connections between the covector mapping theorem and convergence of pseudospectral methods for optimal control. *Computational Optimization and Applications* 41, 307–335.
- Gonzales, R. A., Eisert, J., Koltracht, I., Neumann, M., Rawitscher, G., 1997. Integral equation method for the continuous spectrum radial schrödinger equation. *Journal of Computational Physics* 134 (1), 134–149.
- Gottlieb, D., Hesthaven, J. S., 2001. Spectral methods for hyperbolic problems. *Journal of Computational and Applied Mathematics* 128 (1–2), 83–131.
- Gottlieb, D., Orszag, S. A., 1977. *Numerical Analysis of Spectral Methods: Theory and Applications*. No. 26 in *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.
- Gottlieb, D., Shu, C., 1995a. On the Gibbs phenomenon V: Recovering exponential accuracy from collocation point values of a piecewise analytic function. *Numerische Mathematik* 71, 511–526.
- Gottlieb, D., Shu, C. W., 1995b. On the Gibbs phenomenon IV: Recovering exponential accuracy in a subinterval from a Gegenbauer partial sum of a piecewise analytic function. *Mathematics of Computation* 64 (211), 1081–1095.

- Gottlieb, D., Shu, C.-W., Solomonoff, A., Vandeven, H., November 1992. On the Gibbs phenomenon I: recovering exponential accuracy from the Fourier partial sum of a nonperiodic analytic function. *Journal of Computational and Applied Mathematics* 43 (1–2), 81–98.
- Gottlieb, S., Jung, J., Kim, S., 2011. A review of David Gottlieb’s work on the resolution of the Gibbs phenomenon. *Communications in Computational Physics* 9 (3), 497–519.
- Green, C. D., 1969. *Integral Equation Methods*. Nelson, NY.
- Greengard, L., 1991. Spectral integration and two-point boundary value problems. *SIAM Journal on Numerical Analysis* 28 (4), 1071–1080.
- Greengard, L., Rokhlin, V., 1991. On the numerical solution of two-point boundary value problems. *Communications on Pure and Applied Mathematics* 44 (4), 419–452.
- Grigorenko, I., 2006. *Optimal Control and Forecasting of Complex Dynamical Systems*. World Scientific.
- Guf, J., Jiang, W., 1996. The Haar wavelets operational matrix of integration. *International Journal of Systems Science* 27 (7), 623–628.
- Guo, B. Y., 1998. *Spectral Methods and Their Applications*. World Scientific, Singapore.
- Gustafson, S., Silva, A. R., 1998. On accurate computation of a class of linear functionals. *Journal of Mathematics, Systems, Estimation and Control* 8 (2), 1–12.
- Gustafsson, B., 2011. The work of David Gottlieb: A success story. *Communications in Computational Physics* 9 (3), 481–496.
- Hager, W. W., 2000. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik* 87, 247–282.
- Hargraves, C. R., Paris, S. W., 1987. Direct trajectory optimization using nonlinear programming and collocation. *Journal of Guidance, Control, and Dynamics* 10 (4), 338–342.
- Hendriksen, E., van Rossum, H., 22–27 September 1988. Electrostatic interpretation of zeros. In: *Orthogonal Polynomials and their Applications*. Vol. 1329 of *Lecture Notes in Mathematics*. Springer-Verlag, Segovia, Spain, pp. 241–250.

- Hesthaven, J. S., 1998. Integration preconditioning of pseudospectral operators. I. Basic linear operators. *SIAM Journal on Numerical Analysis* 35 (4), 1571–1593.
- Hesthaven, J. S., May 2000. Spectral penalty methods. *Applied Numerical Mathematics* 33 (1–4), 23–41.
- Hesthaven, J. S., Gottlieb, S., Gottlieb, D., 15 January 2007. *Spectral Methods for Time-Dependent Problems*. Vol. 21 of Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.
- Hicks, G. A., Ray, W. H., 1971. Approximation methods for optimal control synthesis. *The Canadian Journal of Chemical Engineering* 49 (4), 522–528.
- Horadam, A. F., 1985. Gegenbauer polynomials revisited. *The Fibonacci Quarterly* 23, 294–299.
- Hornig, I., Ho, S., 1985. Optimal control using discrete Laguerre polynomials. *International Journal of Control* 41 (6), 1613–1619.
- Hosseini, M. M., 15 May 2006. A modified pseudospectral method for numerical solution of ordinary differential equations systems. *Applied Mathematics and Computation* 176 (2), 470–475.
- Hsiao, C., December 1997. State analysis of linear time delayed systems via Haar wavelets. *Mathematics and Computers in Simulation* 44 (5), 457–470.
- Hsu, N., Cheng, B., 1981. Analysis and optimal control of time-varying linear systems via block-pulse functions. *International Journal of Control* 33 (6), 1107–1122.
- Hua, G., Wang, W., Zhao, Y., Li, L., 2011. A study of an optimal smoke control strategy for an Urban Traffic Link Tunnel fire. *Tunnelling and Underground Space Technology* 26 (2), 336–344.
- Hull, D., 1997. Conversion of optimal control problems into parameter optimization problems. *Journal of Guidance, Control, and Dynamics* 20 (1), 57–60.
- Huntington, G. T., June 2007. *Advancement and Analysis of a Gauss Pseudospectral Transcription for Optimal Control*. Ph.D. dissertation, Department of Aeronautics and Astronautics. Massachusetts Institute of Technology.
- Hwang, C., Chen, M., 1985. Analysis and optimal control of time-varying linear systems via shifted Legendre polynomials. *International Journal of Control* 41 (5), 1317–1330.

- Hwang, C., Shih, D., Kung, F., August 1986. Use of block-pulse functions in the optimal control of deterministic systems. *International Journal of Control* 44 (2), 343–349.
- Hwang, C., Shih, Y. P., 1983. Laguerre series direct method for variational problems. *Journal of Optimization Theory and Applications* 39, 143–149.
- Imani, A., Aminataei, A., Imani, A., 2011. Collocation method via Jacobi polynomials for solving nonlinear ordinary differential equations. *International Journal of Mathematics and Mathematical Sciences* 2011, 1–11.
- Isaacson, E., Keller, H. B., 1994. *Analysis of Numerical Methods*. Dover Books on Mathematics. Dover Publications.
- Jackiewicz, Z., 2003. Determination of optimal parameters for the Chebyshev-Gegenbauer reconstruction method. *SIAM Journal on Scientific Computing* 25 (4), 1187–1198.
- Jackiewicz, Z., Park, R., June 2009. A strategy for choosing Gegenbauer reconstruction parameters for numerical stability. *Applied Mathematics and Computation* 212 (2), 418–434.
- Jaddu, H., 7 January 2002. Spectral method for constrained linear-quadratic optimal control. *Mathematics and Computers in Simulation* 58 (2), 159–169.
- Jaddu, H., Shimemura, E., 1999. Computation of optimal control trajectories using Chebyshev polynomials: parameterization, and quadratic programming. *Optimal Control Applications and Methods* 20 (1), 21–42.
- Jung, J.-H., Gottlieb, S., Kim, S. O., Bresten, C. L., Higgs, D., 2010. Recovery of high order accuracy in radial basis function approximations of discontinuous problems. *Journal of Scientific Computing* 45, 359–381.
- Kadalbajoo, M. K., Yadaw, A. S., July 2008. B-spline collocation method for a two-parameter singularly perturbed convection diffusion boundary value problems. *Applied Mathematics and Computation* 201 (1–2), 504–513.
- Kameswaran, S., Biegler, L., 2008. Convergence rates for direct transcription of optimal control problems using collocation at Radau points. *Computational Optimization and Applications* 41, 81–126.
- Kang, W., December 2008. The rate of convergence for a pseudospectral optimal control method. In: *47th IEEE Conference on Decision and Control*. CDC 2008. pp. 521–527.

- Kang, W., April 2009. On the rate of convergence for the pseudospectral optimal control of feedback linearizable systems. arXiv:0904.0833v1 [math.OC], 1–28.
- Kang, W., 2010. Rate of convergence for the Legendre pseudospectral optimal control of feedback linearizable systems. *Journal of Control Theory and Applications* 8, 391–405.
- Kang, W., Bedrossian, N., September 2007. Pseudospectral optimal control theory makes debut flight, saves NASA \$1M in under three hours. *SIAM News* 40 (7).
- Kang, W., Gong, Q., Ross, I., 12–15 December 2005. Convergence of pseudospectral methods for a class of discontinuous optimal control. In: 44th IEEE Conference on Decision and Control and 2005 European Control Conference. pp. 2799–2804.
- Kang, W., Gong, Q., Ross, I. M., Fahroo, F., 2007. On the convergence of nonlinear optimal control using pseudospectral methods for feedback linearizable systems. *International Journal of Robust and Nonlinear Control* 17 (14), 1251–1277.
- Kang, W., Ross, I. M., Gong, Q., 2008. Pseudospectral optimal control and its convergence theorems. In: Astolfi, A., Marconi, L. (Eds.), *Analysis and Design of Nonlinear Control Systems*. Springer Berlin Heidelberg, pp. 109–124.
- Kanwal, R. P., 1997. *Linear Integral Equations: Theory and Technique*. Birkhauser Boston Academic Press.
- Kaya, C., Martínez, J., August 2007. Euler discretization and inexact restoration for optimal control. *Journal of Optimization Theory and Applications* 134 (2), 191–206.
- Kaya, C. Y., 2010. Inexact restoration for Runge-Kutta discretization of optimal control problems. *SIAM Journal on Numerical Analysis* 48 (4), 1492–1517.
- Keiner, J., 2009. Computing with expansions in Gegenbauer polynomials. *SIAM Journal on Scientific Computing* 31 (3), 2151–2171.
- Kekkeris, G. T., Paraskevopoulos, P. N., 1988. Hermite series approach to optimal control. *International Journal of Control* 47 (2), 557–567.
- Keyser, J., Krishnan, S., Manocha, D., Culver, T., 1998. Efficient and reliable computation with algebraic numbers for geometric algorithms. Technical report tr98-012, Department of Computer Science, University of North Carolina. URL <http://europepmc.org/abstract/CIT/171852>

- Kiparissides, C., Georgiou, A., 1987. Finite-element solution of nonlinear optimal control problems with a quadratic performance index. *Computers & Chemical Engineering* 11 (1), 77–81.
- Kirk, D. E., 2004. *Optimal Control Theory: An Introduction*. Dover books on engineering. Dover Publications.
- Kogut, P. I., Leugering, G. R., 2011. *Optimal Control Problems for Partial Differential Equations on Reticulated Domains: Approximation and Asymptotic Analysis*. Systems and Control. Birkhäuser Boston.
- Kong, W. Y., Rokhlin, V., 2012. A new class of highly accurate differentiation schemes based on the prolate spheroidal wave functions. *Applied and Computational Harmonic Analysis* 33 (2), 226–260.
- Kopriva, D. A., 2009. *Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers*. Springer, Berlin.
- Kotikov, A. V., February 2001. The Gegenbauer polynomial technique: the evaluation of complicated Feynman integrals. arXiv:hep-ph/0102177.
- Krishnan, S., Manocha, D., 1995. Numeric-symbolic algorithms for evaluating one-dimensional algebraic sets. In: *Proceedings of the 1995 international symposium on Symbolic and algebraic computation*. ACM, pp. 59–67.
- Lampe, B., Kramer, G., 1983. Application of Gegenbauer integration method to  $e^+e^-$  annihilation process. *Physica Scripta* 28, 585–592.
- Lang, F., Xu, X., April 2012. Quintic B-spline collocation method for second order mixed boundary value problem. *Computer Physics Communications* 183 (4), 913–921.
- Lashari, A. A., Zaman, G., 2012. Optimal control of a vector borne disease with horizontal transmission. *Nonlinear Analysis: Real World Applications* 13 (1), 203–212.
- Lawton, J. R., Beard, R. W., Mclain, T. W., 1999. Successive Galerkin approximation of nonlinear optimal attitude. In: *Proceedings of the 1999 American Control Conference*. Vol. 6. pp. 4373–4377.
- Lee, T., Tsay, S., 1989. Analysis, parameter identification and optimal control of time-varying systems via general orthogonal polynomials. *International Journal of Systems Science* 20 (8), 1451–1465.

- Lenhart, S., Workman, J., 2007. *Optimal Control Applied to Biological Models*. Chapman and Hall/CRC mathematical & computational biology series. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton.
- Lewis, F. L., Syrmos, V. L., 1995. *Optimal Control*. A Wiley-Interscience publication. J. Wiley.
- Li, Y., Santosa, F., June 1996. A computational algorithm for minimizing total variation in image restoration. *IEEE Transactions on Image Processing* 5 (6), 987–995.
- Light, W. A., 1978. A comparison between Chebyshev and ultraspherical expansions. *Journal of the Institute of Mathematics and its Applications* 21, 455–460.
- Lin, P., Su, S., Lee, T., October 2004. Time-optimal control via fuzzy approach. In: *IEEE International Conference on Systems, Man and Cybernetics*. Vol. 4. pp. 3817–3821.
- Lin, W.-S., 2011. Optimality and convergence of adaptive optimal control by reinforcement synthesis. *Automatica* 47 (5), 1047–1052.
- Liu, B., 2011. Initial condition of costate in linear optimal control using convex analysis. *Automatica* 47 (4), 748–753.
- Liu, C., Shih, Y., 1983. Analysis and optimal control of time-varying systems via Chebyshev polynomials. *International Journal of Control* 38 (5), 1003–1012.
- Liu, D., Gibaru, O., Perruquetti, W., 2011. Differentiation by integration with Jacobi polynomials. *Journal of Computational and Applied Mathematics* 235 (9), 3015–3032.
- Liu, W., Ma, H., Tang, T., Yan, N., 2004. A posteriori error estimates for discontinuous Galerkin time-stepping method for optimal control problems governed by parabolic equations. *SIAM Journal on Numerical Analysis* 42 (3), 1032–1061.
- Liu, W., Yan, N., 2001. A posteriori error estimates for distributed convex optimal control problems. *Advances in Computational Mathematics* 15 (1–4), 285–309.
- Long, G., Sahani, M. M., Nelakanti, G., 1 September 2009. Polynomially based multi-projection methods for Fredholm integral equations of the second kind. *Applied Mathematics and Computation* 215 (1), 147–155.
- Ludlow, I. K., Everitt, J., March 1995. Application of Gegenbauer analysis to light scattering from spheres: Theory. *Physical Review E* 51, 2516–2526.

- Lundbladh, A., Henningson, D. S., Johansson, A. V., 1992. An efficient spectral integration method for the solution of the Navier Stokes equations. Technical Report FFA TN 1992-28, Aeronautical Research Institute of Sweden, Bromma.
- Lurati, L., March 2007. Padé-Gegenbauer suppression of Runge phenomenon in the diagonal limit of Gegenbauer approximations. *Journal of Computational Physics* 222 (1).
- Ma, H., Qin, T., Zhang, W., March 2011. An efficient Chebyshev algorithm for the solution of optimal control problems. *IEEE Transactions on Automatic Control* 56 (3), 675–680.
- Mai-Duy, N., See, H., Tran-Cong, T., 2008. An integral-collocation-based fictitious-domain technique for solving elliptic problems. *Communications in Numerical Methods in Engineering* 24 (11), 1291–1314.
- Mai-Duy, N., Tanner, R. I., 1 April 2007. A spectral collocation method based on integrated Chebyshev polynomials for two-dimensional biharmonic boundary-value problems. *Journal of Computational and Applied Mathematics* 201 (1), 30–47.
- Malek, A., Phillips, T. N., 1995. Pseudospectral collocation method for fourth-order differential equations. *IMA Journal of Numerical Analysis* 15 (4), 523–553.
- Maleknejad, K., Attary, M., July 2011. An efficient numerical approximation for the linear class of Fredholm integro-differential equations based on Cattani’s method. *Communications in Nonlinear Science and Numerical Simulation* 16 (7), 2672–2679.
- Marzban, H. R., Razzaghi, M., June 2003. Hybrid functions approach for linearly constrained quadratic optimal control problems. *Applied Mathematical Modelling* 27 (6), 471–485.
- Marzban, H. R., Razzaghi, M., January 2010. Rationalized Haar approach for nonlinear constrained optimal control problems. *Applied Mathematical Modelling* 34 (1), 174–183.
- Mashayekhi, S., Ordokhani, Y., Razzaghi, M., 2012. Hybrid functions approach for nonlinear constrained optimal control problems. *Communications in Nonlinear Science and Numerical Simulation* 17 (4), 1831–1843.
- Mason, J. C., Handscomb, D. C., 2003. *Chebyshev Polynomials*. Chapman & Hall/CRC, Boca Raton.

- Mercier, B., 1989. An Introduction to the Numerical Analysis of Spectral Methods. Vol. 318 of Lecture Notes in Physics. Springer-Verlag, Berlin; NY.
- Mihaila, B., Mihaila, I., 2002. Numerical approximations using Chebyshev polynomial expansions: El-gendi's method revisited. *Journal of Physics A: Mathematical and General* 35 (3), 731.
- Miller, B., Stepanyan, K., Miller, A., Andreev, M., 12–15 December 2011. 3D Path Planning in a Threat Environment. In: 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC). Orlando, FL, USA, pp. 6864–6869.
- Morrison, D. D., Riley, J. D., Zancanaro, J. F., December 1962. Multiple shooting method for two-point boundary value problems. *Communications of the ACM* 5 (12), 613–614.
- Murray, J., Cox, C. J., Lendaris, G. G., Saeks, R., May 2002. Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 32 (2), 140–153.
- Muskhelishvili, N. I., 1953. *Singular Integral Equations*. Noordhoff, Leiden.
- Nagurka, M. L., Wang, S. K., 1993. A Chebyshev-based state representation for linear quadratic optimal control. *Journal of Dynamic Systems, Measurement, and Control* 115 (1), 1–6.
- Nagurka, M. L., Yen, V., 1990. Fourier-based optimal control of nonlinear dynamic systems. *Journal of Dynamic Systems, Measurement, and Control* 112 (1), 17–26.
- Oh, S. H., Luus, R., 1977. Use of orthogonal collocation method in optimal control problems. *International Journal of Control* 26 (5), 657–673.
- Okosun, K. O., Ouifki, R., Marcus, N., 2011. Optimal control analysis of a malaria disease transmission model that includes treatment and vaccination with waning immunity. *Biosystems* 106 (2–3), 136–145.
- Orszag, S. A., 1971. Accurate solution of the Orr-Sommerfeld stability equation. *Journal of Fluid Mechanics* 50 (4), 659–703.
- Padhi, R., Unnikrishnan, N., Wang, X., Balakrishnan, S. N., December 2006. A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems. *Neural Networks* 19 (10), 1648–1660.

- Paengjuntuek, W., Thanasinthana, L., Arpornwichanop, A., 2012. Neural network-based optimal control of a batch crystallizer. *Neurocomputing* 83 (0), 158–164.
- Palanisamy, K. R., Prasada, R. G., November 1983. Optimal control of linear systems with delays in state and control via Walsh functions. *Control Theory and Applications, IEE Proceedings D* 130 (6), 300–312.
- Paraskevopoulos, P. N., 1983. Chebyshev series approach to system identification, analysis and optimal control. *Journal of the Franklin Institute* 316 (2), 135–157.
- Paraskevopoulos, P. N., Sparis, P. D., Mouroutsos, S. G., 1985. The Fourier series operational matrix of integration. *International Journal of Systems Science* 16 (2), 171–176.
- Paris, S. W., Hargraves, C. R., 1996. OTIS 3.0 Manual. Boeing Space and Defense Group, Seattle.
- Paris, S. W., Riehl, J. P., Sjaww, W. K., 21–24 August 2006. Enhanced procedures for direct trajectory optimization using nonlinear programming and implicit integration. In: *AIAA/AAS Astrodynamics Specialist Conference and Exhibit*. Keystone, Colorado, pp. 1–19.
- Pesch, H. J., 1994. A practical guide to the solution of real-life optimal control problems. *Control Cybernet.* 23, 7–60.
- Phillips, T. N., Karageorghis, A., June 1990. On the coefficients of integrated expansions of ultraspherical polynomials. *SIAM Journal on Numerical Analysis* 27 (3), 823–830.
- Picart, D., Ainseba, B., Milner, F., 2011. Optimal control problem on insect pest populations. *Applied Mathematics Letters* 24 (7), 1160–1164.
- Polak, E., 1973. An historical survey of computational methods in optimal control. *SIAM Review* 15 (2), 553–584.
- Polak, E., 2011. On the role of optimality functions in numerical optimal control. *Annual Reviews in Control* 35 (2), 247–253.
- Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., Mishchenko, E. F., 1962. *The Mathematical Theory of Optimal Processes*; translated from the Russian by K.N. Tirogoff; English edition edited by L.W. Neustadt. Interscience.

- Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1992. Numerical Recipes in Fortran 77: The Art of Scientific Computing, 2nd Edition. Cambridge University Press.
- Pytlak, R., June 1998. Runge-Kutta based procedure for the optimal control of differential-algebraic equations. *Journal of Optimization Theory and Applications* 97 (3), 675–705.
- Pytlak, R., 1999. Numerical Methods for Optimal Control Problems With State Constraints. No. 1707 in *Lecture Notes in Mathematics*. Springer.
- Quarteroni, A., Valli, A., 1994. Numerical Approximation of Partial Differential Equations, 1st Edition. Vol. 23 of Springer Series in Computational Mathematics. Springer-Verlag, 2nd corr. printing.
- Rao, A., 11–14 August 2003. Extension of a pseudospectral Legendre method to non-sequential multiple-phase optimal control problems. In: *AIAA Guidance, Navigation, and Control Conference and Exhibit*. Austin, TX.
- Rao, A. V., Benson, D. A., Darby, C., Patterson, M. A., Francolin, C., Sanders, I., Huntington, G. T., April 2010. Algorithm 902: GPOPS, A MATLAB software for solving multiple-phase optimal control problems using the Gauss pseudospectral method. *ACM Transactions on Mathematical Software* 37 (2).
- Rao, V., Rao, K., April 1979. Optimal feedback control via block-pulse functions. *IEEE Transactions on Automatic Control* 24 (2), 372–374.
- Razzaghi, M., 1990. Optimal control of linear time-varying systems via Fourier series. *Journal of Optimization Theory and Applications* 65, 375–384.
- Razzaghi, M., Elnagar, G., 1993. A Legendre technique for solving time-varying linear quadratic optimal control problems. *Journal of the Franklin Institute* 330 (3), 453–463.
- Razzaghi, M., Elnagar, G. N., 1994. Linear quadratic optimal control problems via shifted Legendre state parametrization. *International Journal of Systems Science* 25 (2), 393–399.
- Razzaghi, M., Razzaghi, M., 1990. Solution of linear two-point boundary value problems and optimal control of time-varying systems by shifted Chebyshev approximations. *Journal of the Franklin Institute* 327 (2), 321–328.
- Razzaghi, M., Razzaghi, M., Arabshahi, A., 1990. Solutions of convolution integral and Fredholm integral equations via double Fourier series. *Applied Mathematics and Computation* 40 (3), 215–224.

- Razzaghi, M., Yousefi, S., 2001. The Legendre wavelets operational matrix of integration. *International Journal of Systems Science* 32 (4), 495–502.
- Reddien, G., 1979. Collocation at Gauss points as a discretization in optimal control. *SIAM Journal on Control and Optimization* 17 (2), 298–306.
- Reeger, J., 2009. A comparison of transcription techniques for the optimal control of the International Space Station. M.Sc. thesis, Rice University, Houston, Texas.
- Reid, W. T., 1972. *Riccati Differential Equations*. Mathematics in Science and Engineering. Academic Press.
- Rosenbrock, H. H., Storey, C., 1966. *Computational Techniques for Chemical Engineers*. Pergamon, London.
- Ross, I., Fahroo, F., 2003. Legendre pseudospectral approximations of optimal control problems: New trends in nonlinear dynamics and control and their applications. In: Kang, W., Borges, C., Xiao, M. (Eds.), *Lecture Notes in Control and Information Sciences*. Vol. 295. Springer Berlin/Heidelberg, pp. 327–342.
- Ross, I. M., February 2004. User’s manual for DIDO: A MATLAB application package for solving optimal control problems. Technical report 04-01.0, Tomlab Optimization Inc.
- Ross, I. M., Fahroo, F., 21–26 July 2002. A direct method for solving nonsmooth optimal control problems. In: *Proceedings of the 15th World Congress of the International Federation of Automatic Control*. Barcelona, Spain.
- Ross, I. M., Fahroo, F., 2004. Pseudospectral knotting methods for solving nonsmooth optimal control problems. *Journal of Guidance Control and Dynamics* 27 (3), 397–405.
- Rungger, M., Stursberg, O., 2011. A numerical method for hybrid optimal control based on dynamic programming. *Nonlinear Analysis: Hybrid Systems* 5 (2), 254–274.
- Ruths, J., Zlotnik, A., Li, S., December 2011. Convergence of a pseudospectral method for optimal control of complex dynamical systems. In: *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*. pp. 5553–5558.

- Sager, S., Bock, H., Reinelt, G., 2009. Direct methods with maximal lower bound for mixed-integer optimal control problems. *Mathematical Programming* 118, 109–149, 10.1007/s10107-007-0185-6.
- Salmani, M., Büskens, C., 2011. Real-time control of optimal low-thrust transfer to the Sun-Earth  $L_1$  halo orbit in the bicircular four-body problem. *Acta Astronautica* 69 (9–10), 882–891.
- Schwartz, A., Polak, E., 1996. Consistent approximations for optimal control problems based on Runge-Kutta integration. *SIAM Journal on Control and Optimization* 34 (4), 1235–1269.
- Schwartz, A. L., 1996. *Theory and Implementation of Numerical Methods Based on Runge-Kutta Integration for Solving Optimal Control Problems*. Ph.D. dissertation, University of California, Berkeley.
- Shih, D., Kung, F., May 1986. Optimal control of deterministic systems via shifted Legendre polynomials. *IEEE Transactions on Automatic Control* 31 (5), 451–454.
- Shyu, K., Hwang, C., 1988. Optimal control of linear time-varying discrete systems via discrete Legendre orthogonal polynomials. *Journal of the Franklin Institute* 325 (4), 509–525.
- Singh, R., Pal, B. C., Jabr, R., Lang, P. D., October 2008. Distribution system load flow using primal dual interior point method. In: *Joint International Conference on Power System Technology and IEEE Power India Conference. POWERCON 2008*. pp. 1–5.
- Singh, T., 2010. *Optimal Reference Shaping for Dynamical Systems: Theory and Applications*. CRC Press.
- Sirisena, H., August 1973. Computation of optimal controls using a piecewise polynomial parameterization. *IEEE Transactions on Automatic Control* 18 (4), 409–411.
- Sirisena, H., Tan, K., August 1974. Computation of constrained optimal controls using parameterization techniques. *IEEE Transactions on Automatic Control* 19 (4), 431–433.
- Sirisena, H. R., Chou, F. S., 1981. State parameterization approach to the solution of optimal control problems. *Optimal Control Applications and Methods* 2 (3), 289–298.

- Smith, R. J., 2008. Explicitly accounting for antiretroviral drug uptake in theoretical HIV models predicts long-term failure of protease-only therapy. *Journal of Theoretical Biology* 251 (2), 227–237.
- Srirangarajan, H. R., Srinivasan, P., Dasarathy, B. V., 1975. Ultraspherical polynomials approach to the study of third-order non-linear systems. *Journal of Sound and Vibration* 40 (2), 167–172.
- Stoer, J., Bulirsch, R., 1980. *Introduction to Numerical Analysis*. Springer-Verlag, NY.
- Stryk, O., 1993. Numerical solution of optimal control problems by direct collocation. In: Bulirsch, R., Miele, A., Stoer, J., Well, K. (Eds.), *Optimal Control*. Vol. 111 of ISNM International Series of Numerical Mathematics. Birkhauser Basel, pp. 129–143.
- Subchan, D. S., Zbikowski, R., 2009. *Computational Optimal Control: Tools and Practice*. John Wiley & Sons.
- Süli, E., Mayers, D. F., 2003. *An Introduction to Numerical Analysis*. Cambridge University Press.
- Szegö, G., 1975. *Orthogonal Polynomials*, 4th Edition. Vol. 23. American Mathematical Society Colloquium Publications.
- Tang, T., Trummer, M. R., 1996. Boundary layer resolving pseudospectral methods for singular perturbation problems. *SIAM Journal on Scientific Computing* 17 (2), 430–438.
- Tang, T., Xu, X., Cheng, J., 2008. On spectral methods for Volterra integral equations and the convergence analysis. *Journal of Computational and Applied Mathematics* 26 (6), 825–837.
- Teo, K. L., Womersley, R. S., 1983. A control parametrization algorithm for optimal control problems involving linear systems and linear terminal inequality constraints. *Numerical Functional Analysis and Optimization* 6 (3), 291–313.
- Teo, K. L., Wong, K. H., 1992. Nonlinearly constrained optimal control problems. *The ANZIAM Journal* 33 (04), 517–530.
- Tian, H., 1989. *Spectral Methods for Volterra Integral Equations*. M.Sc. thesis, Harbin Institute of Technology, Harbin, P.R. China.

- Trefethen, L. N., 21–24 March 1988. Lax-stability vs. eigenvalue stability of spectral methods. In: Numerical methods for fluid dynamics III. Oxford, England; United Kingdom, pp. 237–253.
- Trefethen, L. N., 1996. Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations. Cornell University, [Department of Computer Science and Center for Applied Mathematics].
- Trefethen, L. N., 2000. Spectral Methods in MATLAB. SIAM, Philadelphia.
- Trefethen, L. N., Trummer, M. R., 1987. An instability phenomenon in spectral methods. *SIAM Journal on Numerical Analysis* 24 (5), 1008–1023.
- Tsang, T. H., Himmelblau, D. M., Edgar, T. F., 1975. Optimal control via collocation and non-linear programming. *International Journal of Control* 21 (5), 763–768.
- Tsay, S., Lee, T., 1987. Analysis and optimal control of linear time-varying systems via general orthogonal polynomials. *International Journal of Systems Science* 18 (8), 1579–1594.
- Verhelst, C., Logist, F., Impe, J. V., Helsen, L., 2012. Study of the optimal control problem formulation for modulating air-to-water heat pumps connected to a residential floor heating system. *Energy and Buildings* 45, 43–53.
- Villadsen, J., Stewart, W., 1995. Solution of boundary-value problems by orthogonal collocation. *Chemical Engineering Science* 50 (24), 3981–3996.
- Vlassenbroeck, J., 1988. A Chebyshev polynomial method for optimal control with state constraints. *Automatica* 24 (4), 499–506.
- Vlassenbroeck, J., Dooren, R. V., 1988. A Chebyshev technique for solving nonlinear optimal control problems. *IEEE Transactions on Automatic Control* 33 (4), 333–340.
- von Stryk, O., Bulirsch, R., 1992. Direct and indirect methods for trajectory optimization. *Annals of Operations Research* 37 (1), 357–373.
- Vozovoi, L., Israeli, M., Averbuch, A., 1996. Analysis and application of the Fourier-Gegenbauer method to stiff differential equations. *SIAM Journal on Numerical Analysis* 33 (5), 1844–1863.
- Vozovoi, L., Weill, A., Israeli, M., 1997. Spectrally accurate solution of non-periodic differential equations by the Fourier-Gegenbauer method. *SIAM Journal on Numerical Analysis* 34 (4), 1451–1471.

- Watanabe, S., 1990. Hilbert spaces of analytic functions and the Gegenbauer polynomials. *Tokyo Journal of Mathematics* 13 (2), 421–427.
- Wei, W., Yin, H.-M., Tang, J., 2012. An optimal control problem for microwave heating. *Nonlinear Analysis* 75 (4), 2024–2036.
- Weideman, J. A. C., Reddy, S. C., 2000. A MATLAB differentiation matrix suite. *ACM Transactions of Mathematical Software* 26 (4), 465–519.
- Weisstein, E. W., 2003. *CRC Concise Encyclopedia of Mathematics*, 2nd Edition. Chapman & Hall/CRC.
- Williams, P., 2004. Jacobi pseudospectral method for solving optimal control problems. *Journal of Guidance, Control, and Dynamics* 27 (2), 293–297.
- Williams, P., 2006. A Gauss-lobatto quadrature method for solving optimal control problems. *Australian and New Zealand Industrial and Applied Mathematics Journal* 47, C101–C115.
- Wong, K. H., Clements, D. J., Teo, K. L., 1985. Optimal control computation for nonlinear time-lag systems. *Journal of Optimization Theory and Applications* 47, 91–107.
- Wu, C., Zhang, H., Fang, T., 2007. Flutter analysis of an airfoil with bounded random parameters in incompressible flow via Gegenbauer polynomial approximation. *Aerospace Science and Technology* 11 (7–8), 518–526.
- Wu, J. L., 2009. A wavelet operational method for solving fractional partial differential equations numerically. *Applied Mathematics and Computation* 214 (1), 31–40.
- Xing, X., Chen, Y., Yi, N., February 2010. Error estimates of mixed finite element methods for quadratic optimal control problems. *Journal of Computational and Applied Mathematics* 233 (8), 1812–1820.
- Yan, H., Fahroo, F., Ross, I. M., 25–27 June 2001. Optimal feedback control laws by Legendre pseudospectral approximations. In: *Proceedings of the American Control Conference*. Vol. 3. Arlington, VA, USA, pp. 2388–2393.
- Yang, C., Chen, C., 1994. Analysis and optimal control of time-varying systems via Fourier series. *International Journal of Systems Science* 25 (11), 1663–1678.
- Yang, G., 2007. Earth-moon trajectory optimization using solar electric propulsion. *Chinese Journal of Aeronautics* 20 (5), 452–463.

- Yang, Y., Xiao, Y., Wang, N., Wu, J., 2012. Optimal control of drug therapy: Melding pharmacokinetics with viral dynamics. *Biosystems* 107 (3), 174–185.
- Yen, V., Nagurka, M., 1991. Linear quadratic optimal control via Fourier-based state parameterization. *Journal of Dynamic Systems, Measurement, and Control* 113 (2), 206–215.
- Yen, V., Nagurka, M., 1992. Optimal control of linearly constrained linear systems via state parametrization. *Optimal Control Applications and Methods* 13 (2), 155–167.
- Yilmazer, A., Kocar, C., February 2008. Ultraspherical-polynomials approximation to the radiative heat transfer in a slab with reflective boundaries. *International Journal of Thermal Sciences* 47 (2), 112–125.
- Zahra, W. K., 1 July 2011. A smooth approximation based on exponential spline solutions for nonlinear fourth order two point boundary value problems. *Applied Mathematics and Computation* 217 (21), 8447–8457.
- Zang, T. A., Wong, Y. S., Hussaini, M. Y., 1982. Spectral multigrid methods for elliptic equations. *Journal of Computational Physics* 48 (3), 485–501.
- Zhang, W., Swinton, S. M., 2012. Optimal control of soybean aphid in the presence of natural enemies and the implied value of their ecosystem services. *Journal of Environmental Management* 96 (1), 7–16.
- Zheng, C. H., Kim, N. W., Cha, S. W., 2012. Optimal control in the power management of fuel cell hybrid vehicles. *International Journal of Hydrogen Energy* 37 (1), 655–663.